Journal of Philosophical Logic (2023) 52:1475–1516
https://doi.org/10.1007/s10992-023-09714-7

A Logical Analysis of Instrumentality Judgments: Means-End Relations in the Context of Experience and Expectations

Kees van Berkel[1] · Tim S. Lyon[2] · Matteo Pascucci[3]

Received: 22 October 2021 / Accepted: 13 June 2023 / Published online: 10 August 2023

Abstract
This article proposes the use of temporal logic for an analysis of instrumentality inspired by the work of G.H. von Wright. The first part of the article contains the philosophical foundations. We discuss von Wright's general theory of agency and his account of instrumentality. Moreover, we propose several refinements to this framework via rigorous definitions of the core notions involved. In the second part, we develop a logical system called Temporal Logic of Action and Expectations (TLAE). The logic is inspired by a fragment of propositional dynamic logic based on indeterministic time. The system is proven to be weakly complete relative to its given semantics. We then employ TLAE to formalise and analyse the instrumentality relations defined in the first part of the paper. Last, we point out philosophical implications and possible extensions of our work.

Keywords Action logic · Instrumentality · Means-end relation · Philosophical logic · Temporal logic · Von Wright

✉ Matteo Pascucci
matteo.pascucci@savba.sk

Kees van Berkel
kees.vanberkel@ruhr-uni-bochum.de

Tim S. Lyon
timothy_stephen.lyon@tu-dresden.de

[1] Ruhr Universität Bochum, Institute for Philosophy II, Universitätsstraße, 150, 44801 Bochum, Germany

[2] Technische Universität Dresden, Institute of Artificial Intelligence, Nöthnitzer Str. 46, 01069 Dresden, Germany

[3] Institute of Philosophy of the Slovak Academy of Sciences, v.v.i, Klemensova 19, 811 09 Bratislava, Slovak Republic

1 Introduction

Agents shape their world by making choices, performing actions, and exercising abilities. They reason practically about attaining ends, plan short-term and long-term, and comply with and violate norms. Agents may be right, lucky, and mistaken in their planning. In all of the above, instrumentality statements—also referred to as means-end relations—play a vital role. They provide reasons to act in one way instead of the other. For instance, the statement 'Taking the A-train is an excellent means for going to Harlem' may influence whether I visit my friend Edward, who lives in Harlem, by train. My experience with traffic jams in New York may cause me to refrain from going by car instead.

The concepts of action, ability, and choice are central to any theory of agency [2, 17, 19, 20, 35] and have been extensively investigated in philosophical logic [3, 5, 14, 30, 37]. In contrast, the philosophical and logical investigation of instrumentality relations has thus far received comparably little attention in the literature. In particular, the formal study of how agents acquire and comparatively assess the quality of instrumentality judgments remains to be conducted. This is especially noteworthy due to the role of instrumentality statements in the fields of practical reasoning [4, 15, 20, 28, 36, 39], AI planning [26, 27], and linguistics [16, 29, 31]. Although it is common to analyse reasoning *with* such statements, none of the above accounts treat *how* various instrumentality judgments are obtained, compared, and assessed. To our knowledge, Georg Henrik von Wright is the only philosopher who provides such a philosophical, yet brief, account of instrumentality judgments. This article provides a formal account of instrumentality inspired by von Wright's philosophy.

Consider the following practical problem:

> I would like to have this small parcel opened. What should I do? Which action is the best choice for securing my desired end? For instance, would ripping the parcel's cardboard, cutting the tape with a knife, or using a pair of scissors be most suitable for my purpose?

To satisfactorily address the above problem, I must find out which actions are (most) suitable for attaining my goal. Various challenges arise: I need to somehow collect candidate instruments that will satisfactorily resolve my practical problem. Subsequently, none of the instruments at my disposal may be necessary, that is, there are only sufficient means available. How should candidate means be compared? Which must be preferred and chosen? Is there a difference in the quality of these available instruments?

In the aforementioned fields, instrumentality relations are commonly taken as given. For instance, most accounts of practical reasoning limit illustrations of practical inference to cases of necessary means, yielding oversimplified representations [20]. Exceptions are [15, 24, 33] which discuss versions of comparativism as a way to resolve cases in which various sufficient instruments are available. Nonetheless, in all these accounts instrumentality relations are assumed to be present from the outset, bypassing the following questions:

(q1) *How are instrumentality judgments acquired by an agent?*

(q2) *How can the comparative quality be assessed for such judgments?*

This article addresses questions q1 and q2. Concerning the above practical problem, one may respond to these questions in various ways. For instance, one may invoke experience, apply theoretical knowledge, or follow the advice of a trustful adviser. Given my desire to open the parcel, I may recall from personal experience that using a knife or a pair of scissors always sufficed. Then again, I may recall a piece of advice that one should avoid using knives to open parcels since they may damage the contents. Alternatively, I may search the internet for possible ways to open my parcel. In this article, we propose several ways to compare actions as instruments for a given purpose. In particular, we assign a pivotal role to the agent's past in yielding instrumentality judgments.

Three Criteria Guided by questions q1 and q2, any account of instrumentality should be able to address the following three points: First (I), the account must clarify what it means to say that for an agent α, action Δ is an instrument serving purpose φ. The fact that instrumentality is a relative notion becomes clear when we see that different actions may serve different ends for different agents with different abilities. Considering the previous example, I might not be comfortable with knives which means that using a knife to open a parcel would not be a suitable instrument for me (although it might be for someone else).

Second (II), the account must not only provide procedures to determine which instruments are suitable for the purpose at hand, but it must also allow for a comparison of the different instruments collected. The assessment of instrumentality judgments is *axiological* (from the Greek 'axía', for 'value'), i.e., it provides a label expressing a specific value of the instruments at hand; e.g., scissors are *good* instruments for opening parcels, although using your hands would be *better*. Eventually, such axiological judgments concerning instruments may serve as a guide in practical decision-making.

As a third point (III), we observe that one of the typical features of instrumentality relations is that they are not given once and for all. In many cases, such judgments are established via inductive arguments, relying on past experience witnessing the connection between the terms involved. As famously noted by David Hume [23], inductive arguments are affected by a fundamental problem: how are we justified in making inferences from an observed connection in the past to instances of that connection of which we have no experience? For example, in forming instrumentality judgments based on experience, an individual agent can often not collect all relevant past cases that would settle the issue. Even then, future cases may still be different. For this reason, judgments of instrumentality are essentially *defeasible*. For instance, such judgments may need to be revised due to new personal experience, additionally secured information on the past, and newly received data from other agents.

Contributions In this work, we address the above three criteria. We do this by developing a formal account of instrumentality. Our account is grounded in von Wright's theory of agency in general and his analysis of instrumentality in particular. In the work 'The Variety of Goodness' [38], von Wright provides a philosophical discussion of instrumentality by which actions can be judged as 'good instruments' for specific purposes. In order to address (I)-(III), we take von Wright's account as a departure point and extend it where necessary. Those parts rooted in von Wright's account are made explicit throughout this work. We define a logical system called the Temporal Logic of Actions and Expectations (TLAE) that will formally capture the developed theory of instrumentality. We prove the weak completeness of TLAE with respect to a corresponding Kripke-style semantics using a variation of the Fischer-Ladner construction for propositional dynamic logic [18]. Employing TLAE, we provide formalisations of the acquired conceptions of (comparative) instrumentality and discuss the philosophical implications of our formal setting.

Outline Section 2 consists of a brief analysis of von Wright's general theory of agency and his theory of instrumental goodness. We refine von Wright's theory by supplementing criteria for comparing instruments that serve the same purpose. In particular, in Sections 2.2 and 2.3, we address criterion I by providing various notions of instrumentality deduced from an agent's past experience. In Section 2.4, we deal with criterion II by providing different ways of comparing instrumentality judgments, yielding various value judgments. Section 2.5 is devoted to the defeasibility of instrumentality judgments as expressed in criterion III, and additionally contains a refinement of our take on criterion I. In Section 3, we discuss the requirements of our formal language and introduce the logical system TLAE, which is a temporal extension of the 'Logic of Actions and Expectations' (LAE) in [7]. In Section 4, we prove that TLAE is weakly complete. After that, in Section 5, we discuss the philosophical implications of our formalism and address criteria I-III formally: we logically formalise different notions of instrumentality (criterion I), present several semantic notions of comparative instrumentality (criterion II), and discuss the defeasible nature of instrumentality statements (criterion III). In Section 6, we address future work.

2 Agency and Instrumentality

The backbone of our philosophical analysis will be von Wright's general theory of agency, as laid out in [35, 37, 38]. We start with a brief analysis of the theory and refer to [3, 7, 32] for a more extensive discussion. We subsequently provide a philosophical analysis of instrumental goodness and comparative instrumentality. The former is primarily based on von Wright's discussion of instrumental goodness as laid out in [38, pp. 19-40]. In short, instrumental goodness covers the study of judgments concerning how *well* instruments serve their purpose (equivalently, how well means serve their ends). Since von Wright's analysis is sparse—i.e., being a sub-topic of his theory of goodness [38]—we extend his account in two ways: (i) we discuss several refined instrumentality notions and (ii) consider possible definitions of comparing instruments. The theory presented in this section will ground the subsequent formalisation.

2.1 Von Wright's General Theory of Agency

According to von Wright, to act is to interfere with the course of nature [35]. Such interference manifests itself in bringing something about or preventing something from happening. What is brought about or prevented is a particular *state of affairs*, i.e., a partial description of the world such as ‘the parcel is open’. To bring about a result φ means to act “in such a manner that the state of affairs that φ is the result of one’s action” [35, p. 13]. Likewise, prevention of φ indicates that one’s action has succeeded in ensuring $\neg\varphi$.

The above concept of action is founded on the notion of *change*. In fact, for von Wright, any theory of agency and action must presuppose an account of change, e.g., see [3]. A change defines a *transition* from an initial state (i.e., the moment of evaluation) to an end-state (i.e., a subsequent moment) [35]. Such transitions can be either agent-independent (e.g., a moon eclipse) or agent-dependent (e.g., me cutting the tape of my parcel). The agent-dependent setting forces a non-deterministic worldview. That is, to bring something about forces at least the following three elements: the initial state, in which the agent finds herself, the actual end-state (which is the state that emerges after the performed action), and an alternative end-state (which would result from the agent not performing the action).

Von Wright discusses various relations between these three states. By bringing together the above account of change with the twofold distinction of bringing about and prevention, he characterises four types of action: *producing*, *destroying*, *suppressing*, *preserving*. The first two bring about something, whereas the latter two prevent something. (We refer to [3, 7] for a discussion of these action types in a formal setting.) The four action types are characterised in Fig. 1. For instance, at (iii), the act of suppressing p indicates that at the initial state $\neg p$ holds, through the agent’s acting $\neg p$ continues to hold, and if the agent would not have acted p would have come about. Atoms are required in specifying the four action types since such types may become conflated when employing an arbitrary formula φ instead, e.g., producing becomes destroying if φ is $\neg p$.

Von Wright’s reading of the four action types is arguably too strong: the agent’s acting in Fig. 1 ultimately decides the fate of p. In the case of producing, by acting, the agent ensures p, whereas, by not-acting, the agent can ensure $\neg p$. In other words, von Wright’s account takes agency as causally sufficient in both directions, e.g., see [3]. Furthermore, in Fig. 1, there is no distinction made between different kinds of action an agent can perform at w_0.

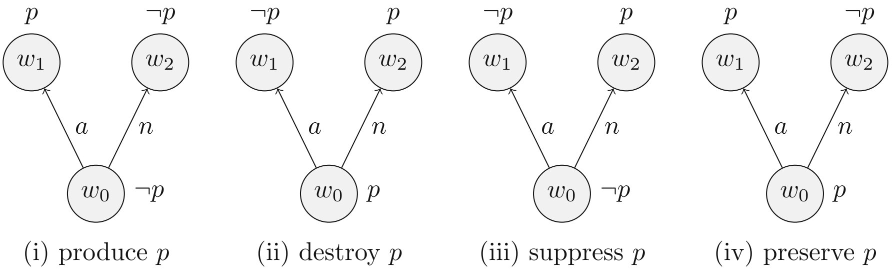

Fig. 1 Von Wright’s four elementary types of action. The transition (arrow) from w_0 denoting the agent acting is labelled a, and the alternative transition indicating the agent’s non-interference with nature is labelled n

Since a general analysis of agency involves many distinct agents and distinct agents can simultaneously perform distinct actions, an individual agent's action is often not causally sufficient. This is known as the *uncertainty* of action [5]. We adopt a generalisation of von Wright's approach that includes this uncertainty. A transition involves the following three elements: (i) an initial state, (ii) a *set* of actions, and (iii) a *set* of possible final states. Henceforth, we also refer to such states as *moments* in indeterministic time. A single agent does not entirely control the course of events, but even the complete set of actions performed by all agents involved does not necessarily entail a unique end-state (cf. the influence of nature). For this reason, we say that a set of actions *causally contributes to the attainment of* the (actual) end-state of a transition only if the end-state would have been different without the performance of that set of actions.

For instance, considering Fig. 2, we say that the action Δ brings the agent from the present moment w_0 to either one of the future moments w_1, w_2, or w_3 without strictly determining either of the three. Still, all three moments satisfy φ. Thus, we say that the agent can bring about φ by performing Δ at w_0, even though the agent cannot secure a unique future moment with action Δ (e.g., one could read φ as 'the parcel is open' and Δ as 'using a knife to cut the parcel's tape'). Furthermore, performing $\overline{\Delta}$—which is the complement of Δ—could lead to at least one future moment where $\neg\varphi$ holds, namely, w_4. Suppose that w_2 is the actual future moment (underlined in Fig. 2). In that case, the agent causally contributed to the attainment of w_2 through performing Δ, since if the agent had not performed Δ, the resulting possible future moment would have been one of $w_4, \ldots, w_n$. Last, although Δ successfully ensures φ, the action can still fail to secure other ends such as ψ in Fig. 2 (e.g., where ψ reads 'the parcel is undamaged').

As a final remark concerning the nature of the term 'action', we will follow the usual distinction between types (i.e., generic categories, such as 'writing') and tokens (i.e., concrete instances in specific circumstances, such as the action of a particular person writing on a particular blackboard on a specific date; see [19] for an extensive

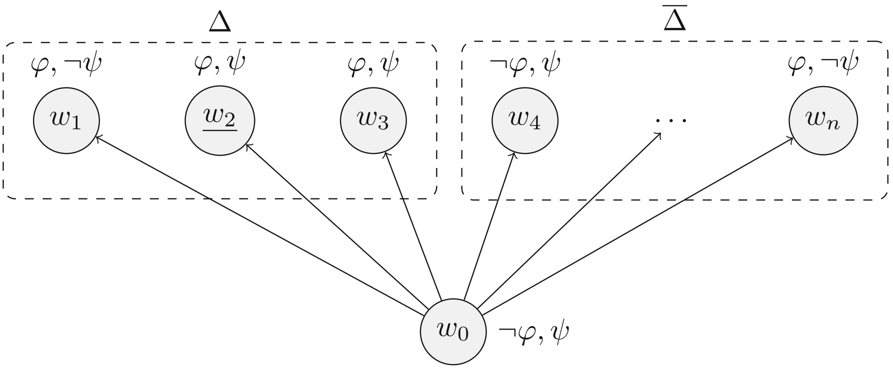

Fig. 2 Indeterministic time and uncertainty of action: bringing about φ through performing action Δ at w_0. Let φ stand for 'the parcel is open', ψ for 'the parcel is undamaged', and Δ for 'using a knife to cut the parcel's tape'. Alternatively, one may label arrows to denote Δ-transitions (cf. Fig. 3)

discussion). Von Wright adopts a similar demarcation by distinguishing between act-categories, on the one hand, and act-individuals, on the other hand [35, p. 36]. It suffices to restrict our analysis to *atomic* actions, *negative* actions (e.g., 'not crossing the street') and *complex* actions (e.g., 'turning left or turning right' and 'turning left and hitting the break'). We do not consider *sequences* of actions.

2.2 Instrumentality and Instrumental Goodness

We are now in a position to address criterion I of Section 1, and clarify what it means that an action Δ is an instrument serving purpose φ. Our analysis will yield two provisional definitions at the end of the next subsection. Central to the study of instrumentality is the relation between an *action* and a *result*. The former is the instrument for the desired outcome expressed in the latter. The outcome can therefore be seen as the purpose of performing the action in question. Thus, we refer to the action as an *instrument* serving a particular *purpose*. For example, 'pulling down the lever of a door and drawing the door towards you' is the instrument for the result 'the door is open'.

Following von Wright [38, p. 21], we can group actions into categories, or kinds, according to the purposes served. To illustrate, no action will qualify as a *cutting*-instrument unless it can serve the purpose of cutting. In this respect, the ability to serve a 'cutting purpose' is a functional characteristic of members of the kind 'cutting'(-instruments). Actions do not necessarily serve a unique purpose. As an instrument, an action can be a member of several kinds, serving multiple purposes. For instance, the action 'using a knife' is an instrument for opening parcels and for peeling apples.

The *goodness* of an instrument is judged in relation to *how* it serves the purpose with which it is associated. Consequently, an instrument may serve certain purposes with excellence while serving others poorly. In this article, we concentrate on a specific agentive source for judgments of instrumental goodness: the agent's (personal) experience with the instruments serving the purpose. In the case of the agent's experience, the agent checks her past for applications of those actions belonging to the same type as the one under consideration, to see whether these past applications successfully served the purpose at hand. This temporal component will be central to our definition of instrumental goodness.

We point out that for instrumentality judgments what matters is an agent's *perception* of the relation between an action and effect. That is, we are concerned with what the agent believes is the relation between an instrument and purpose, which may or may not be rooted in any physical causal relation. For instance, to form a judgement concerning 'cutting knives' and 'opening parcels', an agent needs no (prior) knowledge of the physics involved in cutting the tape of a parcel (the sharpness of the blade, the movement of the arm, the consistency of the tape, etc.). In our case, what counts is the agent's (past) experience of the event (opening a parcel with a knife) and the beliefs that were formed accordingly. Therefore, when we say that a knife is an instrument for opening parcels, we do not refer to its set of physical causal qualities *per se*. In our opinion, the notion of an instrument is a *practical* one, and it would generally be better to keep it distinct from the notion of a cause, as employed in the analysis of *physical* connections between things.[1] We see past experience as a fruitful approach to agentive reasoning since it is a source accessible to the agent at any given time.[2]

This article further develops the idea of rooting instrumentality judgments in past experience. Von Wright involves past experience in his analysis of goodness of skill and instrument, but the expansion of the idea as presented in this article is our own. We propose that instrumentality judgments of this type are formed in *three steps*. In the first step, the agent collects all relevant evidence available from the past. This is the *empirical part* of the procedure. The second step consists in forming context-relative judgments based on what past experience has taught. For instance, in considering all cases in which two actions were performed, it may turn out that a particular outcome was more frequently obtained in association with one of the actions than with the other. This is the *inductive part* of the procedure (i.e., the generalisation of the relevant experience). Third, inferences can be drawn from these context-relative judgments in combination with considerations of chance, (unpredictable) interference, and additional circumstances. Such inferences serve as a guide for the agent's current behaviour. This is the *deductive part* of the argument.

2.3 Three Notions of Instrumentality

The ambivalent nature of the philosophical concept of 'good' may suggest that judgments of instrumental goodness are not objective. However, von Wright [38, p. 25] argues that in dealing with instrumental goodness, the relation between 'purpose', 'instrument' and 'good' is, in fact, objective. Namely, judgments of instrumental goodness express two types of 'connexion'. The first is a causal connexion, expressing a relation between the instrument (as a perceived cause) and its purpose (the desired effect). Von Wright emphasises that this connexion can be empirically checked and is thus objective. Extending the above to cover past experience, we observe that collecting an agent's relevant past experience is also an empirical process, having the same objectivity status. To avoid confusion, we refer to the first connexion as an *empirical connexion* instead. The second is a *logical* connexion that holds between the purpose in question and the term 'better'. Namely, given the agent's desire for a particular effect (the purpose), one can order the available actions according to their empirical connexion such that one action can be logically determined to be a better instrument than another. Accordingly, judgments of instrumental goodness can be objective since we refer to an observable, empirical property of the instrument and assign goodness based on an ordering, which is a judgment of logic.

There are several ways to assess the goodness of an instrument. For von Wright, 'goodness' always refers to particular observable properties of the instrument that make it suitable for some purpose. This property is called the good-making property.

[1] This relates to Hume's criticism of the notion of causation: causality judgments depend on the set of beliefs of a specific group of agents given a specific context and cannot be generalised to constitute physical laws. They are not intended to have such a universal validity.

[2] Past experience is not exhaustive. Additionally, one may consider here-say, advice, education, and theoretical knowledge as sources alternative to past experience. Such alternative sources fulfil an important role in accounting for the employment of instruments that have never been used before. We leave this to future work.

Good-making properties are the properties of the instruments that enable the logical ordering of goodness. To illustrate, if one intends to have some vegetables chopped (the purpose), the good-making property 'sharper' can determine which of the available knives (instruments) is better. Here, sharpness determines the empirical connexion between the knife and cutting. One can then order all available knives according to their sharpness to determine which are sharper and, thus, better [38, p. 25].

This article takes past experience as a general property for ordering instruments. In particular, we focus on the successful applications of the instrument that guaranteed the effect, as witnessed by the agent. Given this good-making property, a judgment about one instrument 'being better' than another is a logical consequence of 'being more successful at guaranteeing the desired result'.

To qualify a generic instrument—i.e., action type—as good for some purpose, we need to ground our judgment in the past performance of concrete actions, i.e., action tokens. We refer to this *temporal component* as the historical witness of an instrument's suitability. In inquiring about whether to apply a certain instrument, agents often base decisions on statements such as:

(i) 'it has worked before';
(ii) 'so far, it has not disappointed me';
(iii) 'well, it has thus far worked better than any of the alternatives'.

These three remarks illustrate the importance of temporal reference. Furthermore, we identify three different criteria of instrumental goodness in (i)-(iii): The first remark exemplifies, what we will call, the minimum criterion for any appropriate definition of instrumentality: (i) the action *has* served the purpose at least once and, for that reason, it *can* serve the purpose as an instrument. That is to say, criterion (i) functions as a lower bound on the instrument's suitability, thus identifying *candidate instruments*. In the second remark, we recognise an upper bound, i.e., a maximum criterion: (ii) there have been applications of the instrument and these applications *have always* served the purpose. Criterion (ii) is referred to as instrumental *excellence*. In the last remark, we identify a *comparative* approach to instrumental goodness: (iii) the action is suitable in *comparison* to alternative actions.

Observe that (i) and (ii) express, respectively, an existential criterion and a universal (constructive) criterion. These criteria enable us to label instruments as 'good', independent of their relation to other instruments. The third (iii) interpretation is a comparative notion that labels instruments as 'good' but only relative to other instruments. If an instrument is not excellent in itself but is the best among others, we refer to the instrument as *comparatively excellent*.

We make our analysis precise in Definition 1. In what follows, we use $\Delta, \Gamma, \Sigma, \ldots$ to denote action types, $\alpha, \alpha_1, \alpha_2, \ldots$ to denote agents, $\varphi, \psi, \chi, \ldots$ to denote propositions (i.e., descriptions of states of affairs), and $w, v, u, \ldots$ to denote moments. The terms 'moment' and 'state' are used interchangeably and the expression 'φ-instrument' abbreviates 'instrument to obtain the state of affairs described by φ'.

Definition 1 (Definition of Candidate Instruments)

(1) CANDIDATE INSTRUMENTS: An action type Δ is a candidate φ-instrument for agent α at moment w if and only if (i) Δ has led to φ for α at least once in the past of w.

(2) EXCELLENT CANDIDATE INSTRUMENTS: An action type Δ is an excellent candidate φ-instrument for agent α at moment w if and only if (i) Δ is a candidate φ-instrument for α at w and (ii) Δ has always led to φ for α in the past of w.

(3) BETTER CANDIDATE INSTRUMENTS: An action type Δ is a better candidate φ-instrument for agent α at moment w than the candidate φ-instruments $\Gamma_1, \ldots, \Gamma_n$ available to α at w if and only if (i) Δ is a candidate φ-instrument for α at w and (ii) in the past of w, action Δ was more successful for α in guaranteeing φ than $\Gamma_1, \ldots, \Gamma_n$.

We point out that judgments of the type expressed above vary over agents, and since the qualification of instruments is defined relative to the past, these judgments are strictly dependent on a vantage point too. This makes the above three definitions *defeasible*: new experience may cause the agent to revise previous judgments (e.g., horses may have been the best option for private transport before cars). We discuss defeasibility in detail in Section 2.5.

2.4 Different Ways of Comparing

In Definition 1, we defined the notion of a 'better instrument' in terms of the *relative success* of the instrument in question. In what follows, we address criterion II of Section 1 and discuss various ways of comparing the success of candidate instruments. The analysis will yield different definitions of comparative instrumental goodness.

Von Wright does not provide an explicit method for comparing candidate instruments, but we find some hints in his analysis of technical goodness (i.e., the goodness of ability, capacity, and skill) [38].[3] There are two ways of testing whether an agent excels at guaranteeing a particular result or at performing some action, namely, (i) through *competition* and (ii) through *achievement* [38, Ch. 2]. Method (i) evaluates personal performance in relation to other agents. In the context of the Olympic games, competition is a means to determine which of all candidate athletes excels at, say, the javelin throw. Such competition can be defined in terms of 'who comes in first' in a single game (the absolute best) or in terms of who ended first in a sequence of games (the overall best). Method (ii) evaluates an agent's performance of an action, not in light of other agents acting, but in relation to a predetermined threshold of excellence. Achievement may be defined as beating a predetermined time or record (which may or may not be previously set by other agents) or achieving a qualifying threshold. Such thresholds function as markers of a specific grade of excellence.

Following von Wright [38, p. 34], both (i) and (ii) are methods of establishing goodness by 'degree of distinction', thus constituting a comparison. Method (i) hints at judging goodness on the basis of ordering the success ratios of participating candidates, whereas method (ii) hints at the usage of pre-established thresholds that suitable candidates must pass.

[3] Von Wright [38] explicitly distinguishes *technical* goodness from *instrumental* goodness. The former relates to agents' skills in obtaining results and applying instruments. The latter denotes an impersonal analysis of how instruments serve purposes. We do not differentiate between instrumental and technical goodness, but the logic developed in this paper does allow for reasoning with agent-dependent and -independent notions of instrumentality.

Translating the above to instrumental goodness, we find two notions of comparison. The first compares a candidate instrument to alternative candidates, which calls for an ordering of (all) suitable instruments serving the purpose at hand. First, we gather all instruments that potentially deserve the label 'good-instrument', these are candidate instruments (see Definition 1). Then, for each candidate, we collect its successful and unsuccessful applications in attaining the desired outcome. Thus, we construct a success-failure ratio for each candidate instrument. An instrument comparatively excels in serving a particular purpose if it has the highest success ratio among its alternatives.

However, we point out that this ordering might fail to properly assess whether an instrumentality relation exists between φ and available actions $\Gamma_1, \ldots, \Gamma_n$. This can happen when some of the considered actions were performed only a small number of times within the considered time. A very small sample size may have no effective statistical power since the observed connections between φ and $\Gamma_1, \ldots, \Gamma_n$ could have resulted from mere chance. In such cases, it is reasonable to set thresholds such that one can avoid engaging in a judgment of those actions that have not been sufficiently tested.

The second method considers a single candidate instrument and checks whether it has satisfied a certain threshold. We identify two types of threshold. The first consists in setting a minimum threshold of the success-failure ratio. For instance, a candidate φ-instrument is a good φ-instrument if it guarantees the effect φ at least half of the time. Although such a threshold is reasonable, there are problems with it. Namely, suppose I throw a dart at the bullseye and end up hitting the bullseye during my first throw (without any practice). Naively, for me throwing darts is an excellent instrument for hitting the bullseye since, thus far, I have always been successful. Often, it makes sense to require an additional threshold in the light of which the candidate instrument must be evaluated. This is the second type: we may require that the instrument has been at least applied n-many times in attempting to attain the end in question (e.g., think of products that must be tested before they can be sold).

Likewise, time plays a role in setting thresholds: imposing a threshold on the number of applications entails imposing a minimum length on the time interval considered. One needs to ensure that the time interval allows for a number of action performances that is at least equal to the threshold. Setting such a limit has another function: it excludes cases that lie too far in the past. For instance, in determining whether I excel at hitting the bullseye it suffices to consider, say, my last 100 attempts. Without a limit, past cases when I was still learning how to play darts would be included, thus misrepresenting my current skills.

Based on the above, we propose the following two definitions of comparison.

Definition 2 (Notions of Comparison) A candidate φ-instrument Δ can be evaluated with respect to:

1. other candidate φ-instruments $\Gamma_1, \ldots, \Gamma_n$ based on a success-failure ratio.
2. a threshold:
   - (i) on a lower bound of past co-occurrences of Δ and φ;
   - (ii) on a lower bound of the ratio of past co-occurrences of Δ and φ versus past occurrences of Δ without φ;

(iii) on both (i) and (ii).

The above distinction gives rise to two types of comparative instrumental goodness: (1) goodness in relation to other instruments and (2) goodness in relation to a set threshold. Combinations of (1) and (2) are also possible: e.g., consider the context of the Olympic games, where an athlete must acquire a certain amount of credits in order to enter the games initially (threshold). An advantage of adopting approach (1) is that it completely preserves the logical status of comparative goodness by ordering available instruments based on their success ratio, independent of any external measurement such as a threshold. The downside of (1) is that it allows for cases such as 'having a lottery ticket is an excellent instrument for winning the lottery since it is the only way to win (despite not being sufficient)'. In such cases, a certain threshold is desirable, such as expressed in approach (2).[4]

Following von Wright, we can label instruments as 'better', 'best', and 'poorest' by ordering the available instruments. The best instrument will be the instrument which is unsurpassed in its success ratio by any other, and the poorest will be unsurpassed in its low success ratio. A poor instrument, however, is different from a(n ethically) bad instrument [38, p. 35]. The former does not serve the purpose well, whereas the latter may serve the purpose well but have (legally or socially) undesirable side effects (e.g., think of 'opening a parcel by first stealing the knife with which you intend to open the parcel').

2.5 Expectations: The Other Temporal Component

So far, we discussed how the past serves as a fruitful source of information for qualifying relations of instrumentality. Judgments of instrumentality are essential for practical reasoning since the latter concerns how states of affairs can be altered through the intervention of an agent (cf. [4, 15, 20, 28]). Practical reasoning is, thus, essentially directed towards the *future*. Through instrumentality judgments, agents may generalise past experience and project it onto the future. This generalisation and projection lies at the heart of the problem of induction. We now address criterion III of Section 1.

In [34], von Wright deals with the problem of induction. He divides the induction problem into two parts (see [34, p. 50]):

(q3) How can we demonstrate that the generalisations we make about experienced cases are correct? How do we know that our generalisation is 'complete'?

(q4) How can we demonstrate that such generalisations are reliable for making predictions? That is, how can we extend our generalisations to the future?

[4] There is another issue concerning the lottery example. Since a myriad of tickets is available for any single lottery draw, and tickets are assigned to buyers randomly, the win-loss proportion should be based on the number of winning tickets vs the number of total tickets. In such cases, reference to an agent is not a primary parameter. The issue relates to the low probability of the instrument serving the purpose. This indicates that some actions may not be suitable as instruments at all. Participating in a lottery n times and winning n times will not make it an excellent instrument. There is no sufficient instrument for winning the lottery (perhaps except by buying all the tickets) since winning is beyond the agent's abilities.

Interestingly, von Wright's division is temporal: (q3) deals with the past and (q4) with the future. Regarding generalisations extending to the past, one can theoretically acquire universally objective judgments by collecting all past instances of the object under generalisation. However, when it comes to predictions, the problem of induction truly shows itself:

> "Scarcely anybody would pretend that predictions, even when based upon the safest inductions, might not fail sometimes." [34, p. 51]

Regarding the future, generalisations are inherently defeasible: future information may falsify our earlier judgments. Although not explicitly stated, von Wright's account of instrumentality judgments (which has an inherently inductive nature) appears to incorporate the same temporal distinction between (i) collecting past cases and (ii) extending past generalisations to the future in terms of predictions. We believe that the discussed theory of instrumentality can account for the problem of induction through the role of *expectations* in instrumentality judgments:

> "Judgments of instrumental goodness, usually, even if not necessarily, contain a conjectural element." [38, p. 27]

In relation to instrumentality, an expectation is a projection of the past onto the nearby future that the action will continue to serve the intended purpose.

In [34], von Wright deals with the general problem of induction as an inherently temporal problem that only arises through the involvement of future cases. Likewise, in the setting of practical reasoning, we find specific roles assigned to the past and the future. We believe that expectation, as a defeasible cognitive attitude, warrants the connection between universal statements about the past and their projection onto the future. Indeed, one can say that even if an agent experienced relations between actions and outcomes in the past, which are not sufficient to form a stable and universally valid judgment about instrumentality, the experience nevertheless serves to support *context-specific and graded* judgments in an agent's practical deliberation. By limiting such judgments to (defeasible) expectations, instead of universal claims, the problem of induction is avoided altogether. Despite being defeasible, such judgments may still guide an agent in decision-making. For instance:

> According to what I have experienced so far, using a knife to cut the tape of a parcel has served the goal of opening a parcel to a sufficiently good degree, and I expect it to work out in the same way, at least in the near future. Using a knife is a good instrument for realising my current goal.

Thus, the limits of inductive arguments do not prevent one from formulating instrumentality judgments to decide how to act in the immediate future.

Now we know that the tentative definitions (Definition 1) provided in Section 2.2 do not suffice, and reference to the conjectural element must be included. In judging whether an instrument is suitable for a present purpose, we take our past experience and

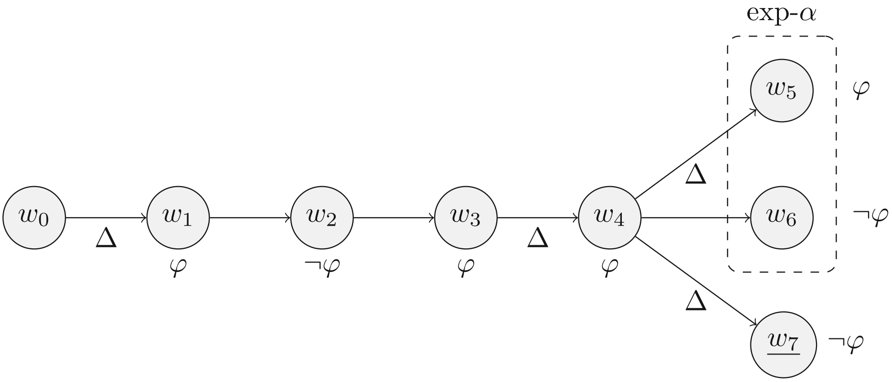

Fig. 3 Expectations and past generalisations of "Δ leading to φ": agent α *expects* (i.e., 'exp-α' and the dotted lines) the moments w_5 and w_6 as possible future continuations of w_4, although w_7 is the actual (underlined) continuation of w_4. The arrows labelled with Δ represent the transitions designated by the performance of Δ

project it onto the future, conjecturing that its past success will sustain in the immediate future. Hence, we recognise two *temporal* components in instrumentality judgments: (i) past performance of particular actions subsumed under a certain type and (ii) the expected continuation of the performance of actions of this type in the nearby future.[5] The first temporal component is related to the empirical part of arguments used to establish instrumentality relations; the second temporal component is related to the inductive part of such arguments (see Section 1).

As a basic example of past generalisations and expectations, consider the model presented in Fig. 3. Let α be an agent at moment w_4. The past (i.e., w_0 up to w_4) provides the agent with the experience that 'so far, φ occurred after every transition caused by Δ'. The agent may generalise this observation to 'generally, performing Δ leads to φ'. Suppose that at w_4, the agent believes that the statement 'Δ leads to φ' will continue to hold in the (nearby) future. The moments w_5 and w_6 are those moments the agent expects to be future continuations of w_4. At w_5 and w_6, it is, in fact, the case that Δ leads to φ. However, the model shows that the actual future continuation of w_4, namely w_7, falsifies the generalisation by making $\neg\varphi$ true after the performance of Δ. This captures the idea that future projections are inherently defeasible: an agent's expectations may turn out to be wrong. She may have expected some other future states to be possible or may have wrongly projected her past experience onto the future. The inclusion of expectations thus allows us to adequately address criterion III.

Based on the above, we extend Definition 1 on candidate instruments to include the conjectural element of instrumentality reasoning. The resulting Definition 3 (below) additionally refines our take on criterion I (the first two items) as well as criterion II (the last two items).

[5] We will see that in the formal framework, the expectations of an agent α correspond to those (nearby) future moments that α regards *likely to happen*. We stress that expectations are not to be confused with epistemic notions of (incomplete) knowledge: an agent can have expectations about the future apart from her knowledge of these expected future moments.

Definition 3 (Four Definitions of Instrument)

(1) INSTRUMENTS: An action type Δ is a φ-instrument for agent α at moment w if and only if (i) Δ has led to φ for α at least once in the past of w and (ii) α expects at w that Δ will lead to φ in the immediate future.

(2) EXCELLENT INSTRUMENTS: An action type Δ is an excellent φ-instrument for agent α at moment w if and only if (i) Δ is a φ-instrument for α at w and (ii) Δ has always led to φ for α in the past of w.

(3) BETTER INSTRUMENTS: An action type Δ is a *better* φ-instrument for agent α at moment w than the φ-instruments $\Gamma_1, \dots, \Gamma_n$ available to α at w if and only if (i) Δ is a φ-instrument for α at w and (ii) in the past of w, action Δ was more successful for α in guaranteeing φ than $\Gamma_1, \dots, \Gamma_n$.

(4) GOOD$_n$ INSTRUMENTS: An action type Δ is a good$_n$ φ-instrument for agent α at moment w if and only if (i) Δ is a φ-instrument for α at w and (ii) α's past performance of Δ satisfies threshold n at w.

Since items (2)-(4) in the above definition are based on Definition 1, all notions of instrumentality are relative to the agent's expectations about the instrument at the moment of evaluation. Items (3) and (4) refer to judgments of instrumentality involving comparisons, respectively thresholds, as specified in Definition 2. The logic of actions and expectations introduced in the subsequent sections allows us to capture these definitions formally.

3 A Temporal Logic of Actions and Expectations: TLAE

On the basis of the analysis provided in Section 2, we develop a logic that enables us to formalise the defined notions of instrumentality. We refer to the logic as the *Temporal Logic of Actions and Expectations*, henceforth TLAE, since these are the main ingredients of its syntax and semantics. TLAE is a linguistic and deductive extension of LAE, a logic developed in [7]. From the linguistic point of view, the novelty is the use of operators for temporal reference to the past. From the deductive point of view, the novelty is the use of axioms that characterise the behaviour of the new operators, as well as their interaction with the old ones.

For the sake of a self-contained exposition, we do not assume familiarity with LAE and provide a detailed presentation of the new framework, mentioning differences with the old one when needed. We start by listing all fundamental concepts that we intend to capture and then introduce the formal language of TLAE in a rigorous way.

3.1 Fundamental Concepts Expressed

Purposes. These are the desired results of actions, represented by formulas $\varphi, \psi, \chi, \dots$ (occasionally annotated). As a matter of fact, a purpose can be equated with a description of a state of affairs. In principle, it is possible to have complex descriptions involving, for instance, modal concepts (e.g., that it possibly rains or it possibly snows). Furthermore, descriptions of states of affairs may also refer to actions (e.g., the description that the door has been opened or the proposition that an agent will open the door).

Actions. These can be (candidate) instruments for achieving a purpose. We use $\delta_1, \delta_2, \delta_3, \ldots$ to represent atomic action types and build complex action types $\Delta, \Gamma, \Theta, \ldots$ (possibly annotated) via complementation '$-$' (overline), union '$\cup$', and intersection '$\cap$' of types (e.g., 'not-opening the door', 'opening the door or opening the window', and 'both opening the door and closing the window', respectively).

Agents. Actions are performed by agents, denoted by $\alpha, \beta, \gamma, \ldots$ (possibly annotated). Different actions may be available to different agents. We will codify the performance of an action type Δ by an agent α as a Boolean formula φ via a function t. Namely, (i) the performance of an atomic action type δ by α is translated via t into a *propositional constant* $\mathfrak{d}^{\alpha}$ and (ii) the performance of a complex action type Δ by α is translated via t into a formula φ whose Boolean structure matches the algebraic structure of Δ (as formally defined below). Notice the difference between the font of, e.g., δ_1, which indicates an *action type*, and the font of $\mathfrak{d}_1^{\alpha_1}$, which indicates a *proposition*, namely that an action of type δ_1 is performed by α_1.

Temporal reference and the structure of time. Judgments of instrumentality refer to the degree in which a certain action, as an instrument, may lead to a certain state of affairs. This 'leading to' is the temporal component of instrumentality referring to possible future moments. We therefore use a modal operator $\Box$, meaning 'in all possible immediate next moments'. This operator is associated with an accessibility relation between pairs of moments, $R_{\Box}$. For instance, let $\mathfrak{d}^{\alpha_1}$ stand for 'the door has been opened by agent α_1', then the formula $\Box\mathfrak{d}^{\alpha_1}$ is interpreted as 'in all possible immediate next moments the door has been opened by agent α_1'. We take it that outcomes of actions concern primarily the nearby future, given that an action affects the world as soon as it is performed by an agent. For this reason, in our framework there is no need of using a temporal operator for reference to the whole future of a moment of evaluation (yet, of course, by iterating operator $\Box$ it is possible to make reference to a more distant future).

We also use two modal operators referring to the past, $\blacksquare$ and H, which extend the language of LAE. The former is the converse of $\Box$ and quantifies over immediate predecessors of a given moment. The latter is the transitive closure of $\blacksquare$, thus referring to the whole past, not just the immediate past. Hence, $\mathsf{H}\varphi$ reads 'everywhere in the past φ holds'. These operators are associated with accessibility relations $R_{\blacksquare}$ and R_{H}, respectively. The main benefit of using operators for past reference in TLAE is that they allow one to *count* previous performances of an action, keeping track of this counting at the syntactic level. Moreover, counting permits us to compute a success ratio with respect to a certain outcome, giving rise to more refined notions of instrumentality. Such a syntactic procedure of counting was not possible in LAE.

Finally, we use an operator $[\mathsf{A}]$ referring to the immediate actual future. This operator, when applied to a formula φ, says that φ describes a state of affairs that takes place in the immediate actual future of the moment of evaluation. This allows us to provide a comparison between what the reasoning agent expects to be the case immediately after the moment of evaluation and what will actually be the case. It is associated with an accessibility relation $R_{[\mathsf{A}]}$ (this operator was originally denoted by N in [7]).

We stress that there is a twofold asymmetry between the past and the future in our formal framework. First, our indeterministic approach to time allows for several possible immediate successors of a moment, but only one immediate predecessor; i.e., the past of a moment is linear, while its future is possibly branching. Second, the way in which instrumentality judgements are formed, namely through past experience, requires reference to the entire past of a moment, but not to its entire future.

Expectations. We employ propositional constants of the form $\mathfrak{e}^{\alpha_i}$ to encode that the most recent expectations of an agent α_i are fulfilled. Such formulae further involve the agent's perspective in judging instrumentality relations. Expectations capture the *conjectural element in judgments of instrumentality.*

3.2 The Formal Language of TLAE

Based on the above list, we define two languages: an action language $\mathcal{L}_{\mathsf{Act}}$, which is an algebra of actions for agent-independent action types, and the logical language $\mathcal{L}_{\mathsf{TLAE}}$ into which these actions will be translated. We let $\texttt{Action}{:=}\{\delta_1, \ldots, \delta_n\}$ be a finite set of *atomic action types* and define the set $\mathcal{L}_{\mathsf{Act}}$ of *all action types* to be all strings generated by the following BNF grammar:

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

where $\delta_i \in \texttt{Action}$. The $\cup$ operation is used to form a *disjunction of action types* (e.g., 'turning-left or turning-right') and the $\overline{}$ operation is used to form a *negation of action types* (e.g., 'not turning-right'). We take the intersection of actions, denoted by the operation $\cap$, as defined, i.e., $\Delta \cap \Gamma{:=}\overline{\overline{\Delta} \cup \overline{\Gamma}}$.

We use $\texttt{Agent}{:=}\{\alpha_1, \ldots, \alpha_m\}$ to denote the set of all agent terms and define an *agent-bound action type* to be an expression of the form Δ^{α_i}, where $\Delta \in \mathcal{L}_{\mathsf{Act}}$ and $\alpha_i \in \texttt{Agent}$. For any $\alpha_i \in \texttt{Agent}$, we let $\texttt{Wit}^{\alpha_i}{:=}\{\mathfrak{d}_1^{\alpha_i}, \ldots, \mathfrak{d}_n^{\alpha_i}\}$ be the set of propositional constants respectively witnessing the performance of action types $\delta_1, \ldots, \delta_n$ by α_i. For instance, suppose δ_1 stands for 'opening the door', then we read its corresponding witness $\mathfrak{d}_1^{\alpha_1}$ as 'the door has been opened by agent α_1'. We make the correspondence between agent-bound action types and propositional constants formally precise below. Notice that $|\texttt{Wit}^{\alpha_i}| = |\texttt{Action}| = n$. We use $\texttt{Wit}$ to denote the set $\bigcup_{\alpha_i \in \texttt{Agent}} \texttt{Wit}^{\alpha_i}$. Last, $\mathfrak{e}^{\alpha_i}$ is a propositional constant witnessing the compatibility of a state with α_i's expectations, which can be read as 'the present state is compatible with α_i's expectations'. $\texttt{Exp} = \{\mathfrak{e}^{\alpha_i} \mid \alpha_i \in \texttt{Agent}\}$ denotes the set of expectation constants for all agents.

The language $\mathcal{L}_{\mathsf{TLAE}}$ of TLAE is defined via the following BNF grammar:

$$\varphi ::= p \mid \mathfrak{e}^{\alpha_i} \mid \mathfrak{d}_j^{\alpha_i} \mid \neg\varphi \mid \varphi \to \varphi \mid \Box\varphi \mid [\mathsf{A}]\varphi \mid \blacksquare\varphi \mid \mathsf{H}\varphi$$

where p is any *propositional variable* from the set $\texttt{Var}{:=}\{p_k \mid k \in \mathbb{N}\}$, $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, n\}$. We use $p, q, r, \ldots$ (possibly annotated) to denote propositional variables, and $\varphi, \psi, \chi, \ldots$ (possibly annotated) to denote formulae from $\mathcal{L}_{\mathsf{TLAE}}$. The connectives $\neg$ and $\to$ denote 'negation' and 'material implication', respectively. The interpretation of each modal operator $\Box$, $\blacksquare$, $[\mathsf{A}]$, and H is given in Definition 5 below; we take $\Diamond$, $\blacklozenge$, $\langle\mathsf{A}\rangle$, and P to be duals of each respective modal operator. Conjunction

$\wedge$, disjunction $\vee$, material equivalence $\leftrightarrow$, *verum* $\top$, and *falsum* $\bot$ are defined in the usual way.

The translation encoding action types from $\mathcal{L}_{\mathsf{Act}}$ into agent-indexed formulae of $\mathcal{L}_{\mathsf{TLAE}}$ is established recursively through the following function t:

- For all $\delta_j \in \texttt{Action}$, and all $\alpha_i \in \texttt{Agent}$, $t(\delta_j^{\alpha_i}) = \mathfrak{d}_j^{\alpha_i}$
- For all $\Delta \in \mathcal{L}_{\mathsf{Act}}$, and all $\alpha_i \in \texttt{Agent}$, $t(\overline{\Delta}^{\alpha_i}) = \neg t(\Delta^{\alpha_i})$
- For all $\Delta, \Gamma \in \mathcal{L}_{\mathsf{Act}}$, and all $\alpha_i \in \texttt{Agent}$, $t(\Delta^{\alpha_i} \cup \Gamma^{\alpha_i}) = t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_i})$

The advantage of this translation is that it enables us to reason with actions on the object language level while simultaneously distinguishing such formulae from other (non-action) formulae in the language. This distinction will prove beneficial in (i) defining a variety of modal instrumentality operators in Section 5 and (ii) axiomatising action specific properties in this section.

To give an example of the expressive power of $\mathcal{L}_{\mathsf{TLAE}}$, we briefly recall the three agentive notions of *would*, *could*, and *will*, as discussed and defined in [7].

$$\begin{aligned}
&Would\\
&[\Delta^{\alpha_i}]^{would}\varphi := \Box(t(\Delta^{\alpha_i}) \rightarrow \varphi). && \text{(d1)}\\
&Could\\
&[\Delta^{\alpha_i}]^{could}\varphi := \Box(t(\Delta^{\alpha_i}) \rightarrow \varphi) \wedge \Diamond t(\Delta^{\alpha_i}). && \text{(d2)}\\
&Will\\
&[\Delta^{\alpha_i}]^{will}\varphi := \Box(t(\Delta^{\alpha_i}) \rightarrow \varphi) \wedge \langle \mathsf{A} \rangle t(\Delta^{\alpha_i}). && \text{(d3)}
\end{aligned}$$

The formula $[\Delta^{\alpha_i}]^{would}\varphi$ (d1) means that 'at the current state, by behaving in accordance with Δ, α_i would bring about φ'. The formula $[\Delta^{\alpha_i}]^{could}\varphi$ (d2) means that 'at the moment of evaluation, 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 $[\Delta^{\alpha_i}]^{will}\varphi$ (d3) means that 'at the moment of evaluation, by behaving in accordance with Δ, α_i would bring about φ and α_i will behave in accordance with Δ'. One can obtain multi-agent variants of the above modalities, such as $[\Delta^{\alpha} \cap \Gamma^{\beta}]^{could}\varphi$, referring to the agents α and β's ability to jointly secure φ. As an example, let Δ be the generic action 'push' and let φ stand for 'the trolley is rolling'. Then, the formula $[\Delta^{\alpha} \cap \Delta^{\beta}]^{will}\varphi$ reads 'α and β will both push to ensure the trolley is rolling'.

There are relevant connections between the modalities 'could', 'would', and 'will' and other operators used in the literature on agency logics. For instance, in the tradition of STIT logics the core ingredients are agent-relative operators which connect an agent's (or a group of agents') behaviour to a certain outcome. The most basic of these operators is $[\alpha\ stit]$, which is extensively analysed in [5]. A formula of the form $[\alpha\ stit]\varphi$ means that agent α behaves in such a way so as to ensure that φ holds. Particularly relevant to our setting is the variant of this operator denoted by $[\alpha\ xstit]$ and analysed in [12, 13], since it involves a temporal shift towards the immediate future: $[\alpha\ xstit]\varphi$ means that agent α behaves in such a way so as to ensure that φ holds *immediately after*. The role played by these operators in STIT logics is here captured by the 'will' modality. As a matter of fact, the formula $[\Delta^{\alpha_i}]^{will}\varphi$ can be regarded as a TLAE-version of $[\alpha\ \ xstit]\varphi$ which additionally includes information about the action chosen by the agent to obtain the result.

There are also significant connections between our approach and proposals to integrate reference to actions in the framework of STIT logics. For instance, the formalism introduced by Xu in [40] includes formulas of the form $[\alpha, \Delta]\varphi$, meaning that agent α obtains φ by doing Δ. This formalism also includes tense-logical operators for future reference, although not for reference to the immediate future. In our framework the situation is reversed: we have operators that make reference to the immediate future and do not have operators that make reference to the whole future (as explained in Section 3.1).

Moreover, as pointed out in [7], the 'would' modality employed here resembles operators used in propositional dynamic logic (PDL) [18]. In a broader perspective, our approach can be seen as a reduction of PDL to alethic modal logic with constants witnessing the performance of actions, similar to the reduction of deontic logic to alethic modal logic proposed by Anderson in [1].[6]

Finally, our account of 'would', 'could', and 'will' is also related to proposals that represent concepts of STIT logic within dynamic logics or logics of temporal computation. For instance, the logic $\mathcal{DLA}$ proposed by Herzig and Lorini [21, 25] is based on a language including an agent-indexed operator $[a : \alpha]$ such that the formula $[a : \alpha]\varphi$ means that agent a ensures that φ will happen by performing an action of type α; this interpretation is very close to our analysis of 'will'. Furthermore, Boudou and Lorini in [10] propose to integrate STIT operators in the computational logic CTL*, whose language is endowed with temporal operators for 'next-time' (X) and 'until' (U). The expressiveness of future reference in the latter setting goes beyond the one allowed for in ours.

3.3 Semantics and Axioms of TLAE

We adopt relational frames for the semantic characterisation of the logic TLAE. The proposed frame properties are motivated by the discussion presented in Section 2. In brief, we define irreflexive tree-like structures that are linear with respect to the past, and allow for branching with respect to the future. Irreflexivity is motivated by the fact that a moment cannot be its own immediate successor. In other words, the transitive closure of the immediate successor relation is a strict partial order. One of the advantages of employing irreflexive structures is that it allows for counting with respect to the past, that is, $\blacklozenge$ and $\blacklozenge\blacklozenge$ refer to two distinct moments in the past, the first immediately preceding the moment of evaluation and the second immediately preceding the first. Henceforth, we use $\blacklozenge^i$ ($i \in \mathbb{N}$) to refer to a concatenation of i-many $\blacklozenge$ operators, referring to a moment i time units in the past. This feature will be employed in counting successful applications of an instrument serving a certain purpose, thus facilitating the comparative notions of instrumentality from Section 2.

[6] One could also add a counterfactual component to the above definitions, namely, the formula $\Diamond\neg\varphi$ as a conjunct. This would strengthen the idea of a causal connection between Δ and φ since it may be the case that φ fails to hold in the immediate future. A counterfactual component is also taken into account in one of Anderson's strategies to reduce deontic logic, as well as in some STIT logics (see, e.g., the deliberative STIT operator in [5]).

First, we define $\mathcal{L}_{\mathsf{TLAE}}$-frames (Definition 4), which will subsequently be refined to form the class of envisaged TLAE-frames (Definition 6). Frames and truth-conditions are defined as usual, with the exception that the interpretation of constants does not vary per model, but is fixed on the level of frames (Definition 5). Namely, the valuation of constants is fixed to sets of moments defined on the frame level, which means that the semantic interpretation of such constants is fixed for every model defined over a frame. One advantage of interpreting constants on the level of frames is that we can define frame properties (and corresponding axioms) that restrict the logical behaviour of certain constants, thus enhancing the expressiveness of the formal setting.

Definition 4 ($\mathcal{L}_{\mathsf{TLAE}}$-frame, $\mathcal{L}_{\mathsf{TLAE}}$-model) A $\mathcal{L}_{\mathsf{TLAE}}$-*frame* is a tuple

$$\mathfrak{F} = \langle W, \{W_{\mathfrak{d}_j^{\alpha_i}} \mid \mathfrak{d}_j^{\alpha_i} \in \mathtt{Wit}\}, \{W_{\mathfrak{e}^{\alpha_i}} \mid \alpha_i \in \mathtt{Agent}\}, R_{\Box}, R_{[\mathsf{A}]}, R_{\blacksquare}, R_{\mathsf{H}}\rangle$$

where W is a set of moments w, u, v, $\ldots$ (which are occasionally annotated), $W_{\mathfrak{d}_j^{\alpha_i}}$, $W_{\mathfrak{e}^{\alpha_i}} \subseteq W$, and $R_{\Box}$, $R_{[\mathsf{A}]}$, $R_{\blacksquare}$, and R_{H} are binary relations over W. We will use alternative (and common) notations to represent the fact that two moments w and u are related by any of these binary relations.

A $\mathcal{L}_{\mathsf{TLAE}}$-*model* $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ is a tuple where $\mathfrak{F}$ is a $\mathcal{L}_{\mathsf{TLAE}}$-frame, and V is a valuation function mapping propositional variables and constants to sets of moments such that:

- $V(\mathfrak{d}_j^{\alpha_i}) := W_{\mathfrak{d}_j^{\alpha_i}}$
- $V(\mathfrak{e}^{\alpha_i}) := W_{\mathfrak{e}^{\alpha_i}}$

Definition 5 (Truth-conditions) Formulae are evaluated at a state of a model, along the following lines:

- $\mathfrak{M}, w \models \chi$ *iff* $w \in V(\chi)$, for any $\chi \in \mathtt{Var} \cup \mathtt{Wit} \cup \mathtt{Exp}$
- $\mathfrak{M}, w \models \neg\varphi$ *iff* $\mathfrak{M}, w \not\models \varphi$
- $\mathfrak{M}, w \models \varphi \rightarrow \psi$ *iff* $\mathfrak{M}, w \not\models \varphi$ or $\mathfrak{M}, w \models \psi$
- $\mathfrak{M}, w \models \Box\varphi$ *iff* for all $v \in W$ s.t. $R_{\Box}wv$, it follows that $\mathfrak{M}, v \models \varphi$
- $\mathfrak{M}, w \models [\mathsf{A}]\varphi$ *iff* for all $v \in W$ s.t. $R_{[\mathsf{A}]}wv$, it follows that $\mathfrak{M}, v \models \varphi$
- $\mathfrak{M}, w \models \blacksquare\varphi$ *iff* for all $v \in W$ s.t. $R_{\blacksquare}wv$, it follows that $\mathfrak{M}, v \models \varphi$
- $\mathfrak{M}, w \models \mathsf{H}\varphi$ *iff* for all $v \in W$ s.t. $R_{H}wv$, it follows that $\mathfrak{M}, v \models \varphi$

The semantic clauses for $\Diamond$, $\langle\mathsf{A}\rangle$, $\blacklozenge$, P, $\wedge$, $\vee$, and $\leftrightarrow$ are defined as usual. We write $\mathfrak{M} \models \varphi$ *iff* for all $w \in W$, $\mathfrak{M}, w \models \varphi$. In this case, we say that φ is valid in $\mathfrak{M}$.

For an arbitrary Δ^{α_i} such that $\Delta \in \mathcal{L}_{\mathsf{Act}}$ and $\alpha_i \in \mathtt{Agent}$, we define $W_{t(\Delta^{\alpha_i})}$ using the following recursive clauses:[7]

[7] Action negation is an extensively debated topic in action logic. Following Broersen [11], there are two main modal approaches to action negation: a universal and a relativised approach. For the former, the negative action modal $\overline{\Delta}$ semantically represents any potential transition between two moments except for those characterised by Δ, i.e., $R_{\overline{\Delta}} := (W \times W) \setminus R_{\Delta}$ (where R_{Δ} expresses a transition between worlds induced by performing Δ). The latter approach defines a negative action modality relative to moments reachable from the moment of evaluation and takes $\overline{\Delta}$ as doing anything but Δ, i.e., $\bigcup_{\Gamma \in \mathcal{L}_{\mathsf{Act}}} R_{\Gamma} \setminus R_{\Delta}$. In [6] an extensive discussion is provided on the relation between the logic 'Logic of Actions and Expectations' (LAE) [7], of which TLAE is an extension, and relativised action negation. In particular, it is shown why the use of action negation in LAE yields the logic compact, whereas the logics in [11] using relativised action negation are not compact.

- $W_{t(\delta_j^{\alpha_i})} := W_{\mathfrak{d}_j^{\alpha_i}}$
- $W_{t(\overline{\Delta}^{\alpha_i})} := W \setminus W_{t(\Delta^{\alpha_i})}$
- $W_{t(\Delta^{\alpha_i} \cup \Gamma^{\alpha_i})} := W_{t(\Delta^{\alpha_i})} \cup W_{t(\Gamma^{\alpha_i})}$

It can be easily shown (through induction on the complexity of Δ^{α_i}) that for each $\Delta \in \mathcal{L}_{\mathsf{Act}}$, each $\alpha_i \in \texttt{Agent}$, each $\mathcal{L}_{\mathsf{TLAE}}$-model $\mathfrak{M}$ and each moment w in its domain, we have:

$$\mathfrak{M}, w \models t(\Delta^{\alpha_i}) \textit{ iff } w \in W_{t(\Delta^{\alpha_i})}$$

Hence, we obtain the following semantic interpretation of our previously defined operator $[\Delta^{\alpha_i}]^{would}\varphi$:

$$\mathfrak{M}, w \models [\Delta^{\alpha_i}]^{would}\varphi \textit{ iff } \text{for all } v \in W \text{ s.t. } R_{\Box}wv, \text{ if } v \in W_{t(\Delta^{\alpha_i})} \text{ then } \mathfrak{M}, v \models \varphi$$

In other words, every immediate successor witnessing the performance of Δ by an agent α_i guarantees the truth of φ.

Definition 6 (TLAE-frame, TLAE-model) We define a TLAE-*frame* to be a $\mathcal{L}_{\mathsf{TLAE}}$-frame satisfying the following properties:

p(A3) For all $w, u, v \in W$, if $R_{[\mathsf{A}]}wu$ and $R_{[\mathsf{A}]}wv$, then $u = v$.

p(A4) For all $w, u \in W$, if $R_{[\mathsf{A}]}wu$, then $R_{\Box}wu$.

p(A5) For all $w \in W$ and all distinct agents $\alpha_1, \ldots, \alpha_n$, if there are (not necessarily distinct) action types $\Delta_1, \ldots, \Delta_n$ s.t. for $1 \leq i \leq n$ there is a $u_i \in W$ s.t. $R_{\Box}wu_i$ and $u_i \in W_{t(\Delta_i^{\alpha_i})}$, there is a $v \in W$ s.t. $R_{\Box}wv$ and $v \in W_{t(\Delta_1^{\alpha_1})} \cap \cdots \cap W_{t(\Delta_n^{\alpha_n})}$.

p(A6) For all $w \in W$ and $\alpha_i \in \texttt{Agent}$, if there is a $v \in W$ s.t. $R_{\Box}wv$ and $v \in W_{\mathfrak{e}^{\alpha_i}}$, then there is also a $u \in W$ s.t. $R_{\Box}wu$ and $u \notin W_{\mathfrak{e}^{\alpha_i}}$.

p(A10;A11) For all $w, v \in W$, $R_{\Box}wv$ iff $R_{\blacksquare}vw$.

p(A12) For all $w, u, v \in W$, if $R_{\blacksquare}wu$ and $R_{\blacksquare}wv$, then $u = v$.

p(A9;A14) R_{H} is the transitive closure of $R_{\blacksquare}$ (that we will occasionally represent as $R_{\blacksquare}^{+}$).

p(A13) For all $w \in W$ either (i) there is no v s.t. $R_{\mathsf{H}}wv$ or (ii) there is some u s.t. $R_{\mathsf{H}}wu$ and there is no z s.t. $R_{\mathsf{H}}uz$.

A TLAE-model $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ is a $\mathcal{L}_{\mathsf{TLAE}}$-model such that $\mathfrak{F}$ is a TLAE-frame. To state that a formula φ is valid in the class of all TLAE-models, we write $\models \varphi$.

For the sake of comparability, we have named the frame properties in reference to the axioms of TLAE introduced below (see Definition 7). Some of the properties presented above correspond to a set of axioms as opposed to a single axiom; e.g., $p(A9;A14)$ is a property characterised through a combination of the axioms $A9$ and $A14$.

Let us briefly explain the intuitive meaning of each property: $p(A3)$ ensures that every moment has at most one immediate actual successor.[8] $p(A4)$ states that the

[8] Item $p(A3)$ expresses the functionality of the operator [A], which is a widespread property among operators used in the literature on agency logic in order to make reference to the actual future. See, e.g., approaches offering a fusion of the next-time operator and STIT operators, such as [12, 21]. In line with [7], we adopt reference to an actual future to discuss various examples in the rest of the article. However, reference to an actual future—i.a., properties $p(A3)$ and $p(A4)$, and axioms $A2$, $A3$, and $A4$, discussed below—can be safely omitted from the technical part of the article (basically, one only has to remove reference to it in the proof of Lemma 4).

actual successor must be a possible successor. $p(A5)$ expresses the agency property known as *independence of agents*.[9] This property ensures that if an agent can perform a certain action at a certain moment, then that agent can perform the action at issue irrespective of the actions performed by the other agents. $p(A6)$ ensures that if an agent expects a certain next moment to arise, then there will be another possible next moment that the agent does not expect to arise. $p(A10;A11)$ defines immediate future moments as the inverse of immediate past moments. $p(A12)$ states that the past is linear. $p(A9;A14)$ defines R_{H} as the transitive closure of $R_{\blacksquare}$. Last, $p(A13)$ ensures that the past is finite, that is, time has a beginning. Consequently, TLAE-frames are rooted tree-like structures. This last property will prove useful to comparing candidate instruments serving the same purpose. We point out the following fact:

Theorem 1 TLAE*-frames are irreflexive, that is, for all* $w \in W$ *of a* TLAE*-frame* $\mathfrak{F}$*, we have* $(w, w) \notin R_{\blacksquare}$, $(w, w) \notin R_{\square}$, $(w, w) \notin R_H$, *and* $(w, w) \notin R_{[\mathsf{A}]}$.

Proof Suppose that $R_{\mathsf{H}}ww$. By $p(A13)$, there is some $u \neq w$ s.t. $R_{\mathsf{H}}wu$ and there does not exist a z such that $R_{\mathsf{H}}uz$. By $p(A9;A14)$ there is a sequence of moments $\sigma = v_1, \ldots, v_n$ s.t. $v_1 = w$, $v_n = u$ and, for $1 \leq i \leq n-1$, $R_{\blacksquare}v_i v_{i+1}$. Furthermore, there is a sequence of moments $\sigma' = v'_1, \ldots, v'_m$ s.t. $v'_1 = v'_m = w$ and for $1 \leq i \leq m-1$, $R_{\blacksquare}v'_i v'_{i+1}$. We now have three cases to consider due to $p(A12)$ and the fact that $v_1 = v'_1$: either (i) σ is a proper sub-sequence of σ', (ii) σ' is a proper sub-sequence of σ, or (iii) $\sigma = \sigma'$. We show (ii) as (i) and (iii) are simple. Since σ' is a sub-sequence of σ, we know that for some $1 \leq i \leq n-1$, $v'_m = w = v_i$. It follows that $R_{\blacksquare}wv_{i+1}$. Moreover, we have that $R_{\blacksquare}wv'_2$ by the definition of σ', which implies that $v_{i+1} = v'_2$ by $p(A12)$. Continuing in this way, one can show that $v_{i+2} = v'_3$, $v_{i+3} = v'_4$, etc. It follows that for some $1 \leq j \leq m$, $v'_j = v_n = u$. However, since σ' forms a cycle, this implies that there is some z (namely, v'_{j+1}) such that $R_{\blacksquare}uv_{j+1}$, which further entails that $R_{\mathsf{H}}uv_{j+1}$ by $p(A9;A14)$. This gives a contradiction.

By $p(A9;A14)$ we can infer that for each w, $(w, w) \notin R_{\blacksquare}$, whence, by $p(A10;A11)$, that $(w, w) \notin R_{\square}$ and finally, by $p(A4)$, that $(w, w) \notin R_{[\mathsf{A}]}$. □

Corollary 1 *The relations* $R_{\square}$, $R_{\blacksquare}$, R_H, *and* $R_{[\mathsf{A}]}$ *of* TLAE*-frames are acyclic.*

Our axiomatisation for the logic TLAE is given below.

Definition 7 (TLAE axiomatisation) The axiomatisation of TLAE consists of the following axiom schemes and rules:

A0 Any propositional tautology
R0 $\varphi, \varphi \rightarrow \psi / \psi$

[9] Independence of agents is a fundamental property in the setting of STIT logic, where it expresses that any choice available to an agent at a certain moment is compatible with any combination of choices available to the other agents at that moment (see Chapter 7 of [5] for a discussion). In STIT, choices are primitive objects and no object-language reference is made to actions. Here, following [7], we adopt independence of agents in a setting in which we have explicit actions available to agents. Such a property ensures that, if an action Δ is a φ-instrument for α, then failing to assure φ by performing Δ cannot be caused by the interference of other agents. Nevertheless, we will discuss possible failures due to the agent's (false) expectations about the action in question. Future work may be directed to investigating instrumentality in the light of interfering agents, i.e., by dropping independence of agents.

A1 $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$

R1 $\varphi/\Box\varphi$

A2 $[\mathsf{A}](\varphi \to \psi) \to ([\mathsf{A}]\varphi \to [\mathsf{A}]\psi)$

A3 $\langle\mathsf{A}\rangle\varphi \to [\mathsf{A}]\varphi$

A4 $\Box\varphi \to [\mathsf{A}]\varphi$

A5 For any distinct $\alpha_1, \ldots, \alpha_n \in \mathtt{Agent}$ and non-necessarily distinct $\Delta_1, \ldots, \Delta_n \in \mathcal{L}_{\mathsf{Act}}$, $(\Diamond t(\Delta_1^{\alpha_1}) \wedge \cdots \wedge \Diamond t(\Delta_n^{\alpha_n})) \to \Diamond(t(\Delta_1^{\alpha_1}) \wedge \cdots \wedge t(\Delta_n^{\alpha_n}))$

A6 For any $\alpha_j \in \mathtt{Agent}$, $\Diamond\mathfrak{e}^{\alpha_j} \to \Diamond\neg\mathfrak{e}^{\alpha_j}$

A7 $\mathsf{H}(\varphi \to \psi) \to (\mathsf{H}\varphi \to \mathsf{H}\psi)$

A8 $\blacksquare(\varphi \to \psi) \to (\blacksquare\varphi \to \blacksquare\psi)$

A9 $\mathsf{H}\varphi \leftrightarrow (\blacksquare\varphi \wedge \blacksquare\mathsf{H}\varphi)$

A10 $\varphi \to \Box\blacklozenge\varphi$

A11 $\varphi \to \blacksquare\Diamond\varphi$

A12 $\blacklozenge\varphi \to \blacksquare\varphi$

A13 $\mathsf{H}\bot \vee \mathsf{PH}\bot$

A14 $\mathsf{H}(\varphi \to \blacksquare\varphi) \to (\blacksquare\varphi \to \mathsf{H}\varphi)$

R2 $\varphi/\mathsf{H}\varphi$

For any formula $\varphi \in \mathcal{L}_{\mathsf{TLAE}}$, we define φ to be a *theorem*, and write $\vdash \varphi$, *iff* (i) φ is an axiom instance, or (ii) φ is derivable from ψ (and χ) via one of the inference rules where $\vdash \psi$ (and $\vdash \chi$, resp.). We say that ψ is *derivable* from Γ in TLAE, written $\Gamma \vdash \psi$, *iff* there exist $\varphi_1, \ldots, \varphi_n \in \Gamma$ s.t. $\vdash \varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$.

In Definition 7 above, we define the notion of a theorem recursively as in [9, Section 4.8], that is, a theorem is a formula which can be derived via a sequence of axiom instances and rule applications to previously derived theorems. We remark that axioms A1, A2, A7 and A8, together with rules R1, R2, and the derivable rules $\varphi/[\mathsf{A}]\varphi$ and $\varphi/\blacksquare\varphi$, qualify the modal operators $\Box$, $[\mathsf{A}]$, $\blacksquare$, and H as normal. Axioms A3 and A12 qualify the accessibility relations associated with $[\mathsf{A}]$ and $\blacksquare$ as functional. Axiom A4 makes the accessibility relation associated with $\Box$ a superset of the accessibility relation associated with $[\mathsf{A}]$. Axiom A5 corresponds to the 'independence of agents' principle in the STIT-literature [5], and states that if each agent can perform a particular action, then all agents can jointly perform such actions. Axioms A9 and A14 are used to express the fact that the accessibility relation associated with H is the transitive closure of the accessibility relation associated with $\blacksquare$. Axioms A10 and A11 are used to express the fact that the accessibility relations associated with $\Box$ and $\blacksquare$ are reciprocally converse. Finally, axiom A13 says that for any state there is a finite sequence of states related to it via the accessibility relation for H. Taken together, these axioms ensure that the accessibility relations for H, $\Box$, $\blacksquare$ and $[\mathsf{A}]$ are irreflexive. However, none of the axioms can ensure this property if taken alone—this follows from general results in correspondence theory for multi-modal, normal logics (cf. [9]). We note that the logic LAE originally presented in [7] is the fragment of TLAE without operators H and $\blacksquare$ and axiomatised with the deductive principles A0-A6 and R0-R1.

3.4 Agentive Modals in TLAE

In order to get an impression of the expressiveness of TLAE, we provide some formal definitions of agentive modals in the spirit of von Wright's analysis. Our basic building blocks will be the agentive modals 'would' (d1), 'could' (d2), and 'will' (d3) defined in Section 3.1. First, consider the four elementary action types as presented in Fig. 1 (recall from Section 2.1 that the four action types require atoms in their formulation):

$$\begin{aligned}&\textit{Produce}\\&[\Delta^{\alpha_i}]^{prod}p{:=}\neg p\wedge[\Delta^{\alpha_i}]^{will}p\wedge\Diamond\neg p \qquad \text{(d4)}\\&\textit{Destroy}\\&[\Delta^{\alpha_i}]^{destr}p{:=}p\wedge[\Delta^{\alpha_i}]^{will}\neg p\wedge\Diamond p \qquad \text{(d5)}\\&\textit{Suppress}\\&[\Delta^{\alpha_i}]^{supp}p{:=}\neg p\wedge[\Delta^{\alpha_i}]^{will}\neg p\wedge\Diamond p \qquad \text{(d6)}\\&\textit{Preserve}\\&[\Delta^{\alpha_i}]^{pres}p{:=}p\wedge[\Delta^{\alpha_i}]^{will}p\wedge\Diamond\neg p \qquad \text{(d7)}\end{aligned}$$

Von Wright's action types are often referred to as *deliberative* in nature; that is, they exclude outcomes which are trivial (e.g., $\top$) and ensure that outcomes are about contingent states of affairs p, namely, for which $\Diamond p$ and $\Diamond\neg p$ hold. As discussed in Section 2, von Wright's reading of these actions may be too strong, namely, the agent's action decides the faith of p completely. For instance, in the case of 'producing', through acting the agent ensures p whereas through not-acting the agent is able to ensure $\neg p$. In other words, von Wright's account takes the agent's agency as causally sufficient in both directions. In line with our discussion, taking a slightly weaker standpoint (cf. *causal contribution* in Fig. 2), we alternatively formalise that the agent has the ability to bring about p through performing Δ, but does not bring about p by not performing Δ. This is reflected in Definitions (d4), (d5), (d6), and (d7). (We observe that this position was already adopted in [3].)

By making use of the notions of 'would' (d1) and 'could' (d2), one can provide new versions of the four action types presented above. We call such variations *volitional* concepts. For instance, (d8) expresses the idea that an agent α_i could destroy p by performing the action Δ. In particular, the first conjunct of (d8) states that p is presently the case, the second ensures that by performing Δ the agent α_i would bring about $\neg p$ and Δ can be performed by α_i, and last it is possible that p will not be destroyed.

$$\begin{aligned}&\textit{Could Destroy}\\&[\Delta^{\alpha_i}]^{could}_{destr}p{:=}p\wedge[\Delta^{\alpha_i}]^{could}\neg p\wedge\Diamond p \qquad \text{(d8)}\end{aligned}$$

Last, consider the notion of *forbearance*, which is a stronger agentive notion than merely not acting according to von Wright. To be more precise, forbearing assumes the agent's *ability* to perform the action that is forborne.[10] The formal definition of forbearance (irrespective of its outcome, denoted by '$\top$') is presented in (d9).

$$\begin{array}{l} Forbear \\ [\Delta^{\alpha_i}]^{forb}\top := [\Delta^{\alpha_i}]^{could}\top \wedge [\overline{\Delta}^{\alpha_i}]^{will}\top \end{array} \tag{d9}$$

In words, (d9) reads 'the agent α_i forbears performing action Δ whenever α_i could perform action Δ, but will instead perform the action's complement $\overline{\Delta}$'. One can see how the notion of forbearance can be extended to incorporate the four elementary action types. The definition provided in (d10) gives an example.

$$\begin{array}{l} Forbear\ to\ Produce \\ [\Delta^{\alpha_i}]^{forb}_{prod}p := p \wedge [\Delta^{\alpha_i}]^{could}p \wedge [\overline{\Delta}^{\alpha_i}]^{will}\top \wedge \Diamond\neg p \end{array} \tag{d10}$$

So far, we only considered temporal operators referring to the future. We briefly point out that the modularity of combining complex modals of agency extends to reasoning about the past. For instance, one can combine the four elementary action types and the notion of forbearance, with the three notions of 'would', 'could', and 'will', while referring to the agent's past. We investigate reasoning about the past when we formalise instrumentality in Section 5.

The aim of the above discussion is to demonstrate the high versatility in defining formal notions of agency in the language of TLAE. One of the reasons for this expressiveness relates to the use of action constants. Namely, the use of constants referring to actions allows us to distinctively reason about actions and states of affairs in a highly modular way, combining them freely with the available temporal operators: future, past, and actual future. In Section 5, we demonstrate how this language can be employed to express various instrumentality notions. Furthermore, we will consider agentive modals that arise by involving the notion of *expectations*. Before moving to our analysis of instrumentality, we demonstrate that TLAE is consistent, sound, and weakly complete.

4 Soundness and Weak Completeness of TLAE

Due to the interaction between the two new operators for past reference, $\blacksquare$ and H (in particular, the fact that we want one to be the transitive closure of the other), proving the completeness of TLAE requires a much more complex construction than the one

[10] We note that von Wright's concept of forbearing is conceptually different from Belnap et al's [5] notion of refraining. Briefly, von Wright considers 'zero action' or 'passivity' as a meaningful notion, occurring when an agent does not act and lets the course of nature take over (see Fig. 1). This idea can be captured in the definition of forbearing (d9), by taking the negative action $\overline{\Delta}$ to correspond to a conjunction of the negation of each atomic action. This contrasts with Belnap et al's account where refraining from acting necessarily corresponds to the agent actively performing some other action. Nevertheless, in light of STIT, Horty and Belnap [22] argue that von Wright's notion of forbearing is logically equivalent to the deliberative STIT reading of forbearing.

for LAE provided in [7]. As a matter of fact, the usual canonical model construction cannot be used since the logic TLAE is not compact (and hence, not *strongly* complete, cf. [9]): one can prove that the infinite set $\Sigma = \{\blacksquare^n p : n \in \mathbb{N}\} \cup \{\neg \mathsf{H} p\}$ has no TLAE-model, whereas each of its finite subsets has some TLAE-model. The strategy followed in this section consists of adapting the Fischer-Ladner construction for the completeness of propositional dynamic logic (illustrated in [9, Section 4.8]) in order to obtain a weak completeness result for our logic TLAE.

First, TLAE is sound with respect to the class of TLAE-frames:

Theorem 2 *(Soundness) For any formula $\varphi \in \mathcal{L}_{\mathsf{TLAE}}$, if $\vdash \varphi$, then $\models \varphi$.*

Proof Straightforward by demonstrating that all axioms of TLAE are valid for the class of TLAE-frames and all rules of TLAE preserve validity (see [9]). □

Furthermore, we observe that the logic TLAE is consistent.

Theorem 3 *(Consistency) The logic* TLAE *is consistent.*

Proof To show TLAE consistent, we show that the class of models for TLAE is non-empty. We define a TLAE-model $\mathfrak{M}$ as follows: the set of moments $W{:=}\{w_i \mid i \in \mathbb{N}\}$, for each $\mathfrak{d}_j^{\alpha_i} \in \mathtt{Wit}$, $W_{\mathfrak{d}_j^{\alpha_i}}{:=}\emptyset$, for each $\alpha_i \in \mathtt{Agent}$, $W_{\mathfrak{e}^{\alpha_i}}{:=}\emptyset$, $R_{[\mathsf{A}]}{:=}R_{\Box}{:=}\{(w_i, w_{i+1}) \mid i \in \mathbb{N}\}$, $R_{\blacksquare}$ is taken to be the converse of $R_{\Box}$, R_{H} is the transitive closure of $R_{\blacksquare}$, and V is taken to be an arbitrary valuation. It is straightforward to verify that $\mathfrak{M}$ is a TLAE-model. □

We now define a sequence of concepts that will assist us in establishing our weak completeness result.

Definition 8 (TLAE-Closure) Let Σ be a finite set of formulae. The TLAE-closure of Σ is the smallest set $Cl(\Sigma)$ satisfying the conditions below:

- if $\varphi \in \Sigma$ or φ is a subformula of some $\psi \in \Sigma$, then $\varphi \in Cl(\Sigma)$
- if $\mathsf{P}\varphi \in \Sigma$, then $\blacklozenge \mathsf{P}\varphi, \blacklozenge\varphi \in Cl(\Sigma)$
- each constant $\mathfrak{d}_i^{\alpha_j}$ and $\mathfrak{e}^{\alpha_j}$ is in $Cl(\Sigma)$
- if $\Diamond t(\Delta_1^{\alpha_1}), \dots, \Diamond t(\Delta_n^{\alpha_n}) \in \Sigma$, then $\Diamond(t(\Delta_1^{\alpha_1}) \wedge \cdots \wedge t(\Delta_n^{\alpha_n})) \in Cl(\Sigma)$
- if $\Diamond \mathfrak{e}^{\alpha_j} \in \Sigma$, then $\Diamond\neg\mathfrak{e}^{\alpha_j} \in Cl(\Sigma)$
- $\mathsf{H}\bot, \mathsf{PH}\bot \in Cl(\Sigma)$

Definition 9 (Negation $\sim$)

$$\sim\varphi{:=}\begin{cases} \psi & \text{if } \varphi \text{ is of the form } \neg\psi, \\ \neg\varphi & \text{otherwise.} \end{cases}$$

For a given finite set of formulae Σ, we define $\neg Cl(\Sigma)$ to be the smallest extension of $Cl(\Sigma)$ closed under $\sim$ (i.e., under single negations).

Definition 10 (Atomic Set) Let Σ be a finite set of formulae. We say that a set X of formulae is TLAE-*consistent iff* $X \nvdash \bot$, and say that a set X of formulae is *maximally* TLAE-*consistent iff* X is consistent and for any formula $\varphi \notin X$, $X \cup \{\varphi\} \vdash \bot$. A set of formulae X is an *atomic set* over Σ *iff* it is a maximal TLAE-consistent subset of $\neg Cl(\Sigma)$. We use $At(\Sigma)$ to denote the set of all atomic sets over Σ.

Lemma 1 *Let Σ be a finite set of formulae and $X \in At(\Sigma)$. Then,*

(i) For all $\varphi \in \neg Cl(\Sigma)$, either $\varphi \in X$ or $\sim\varphi \in X$, but not both.
(ii) For all $\varphi \in \neg Cl(\Sigma)$, if $X \vdash \varphi$, then $\varphi \in X$.
(iii) For all $\varphi \vee \psi \in \neg Cl(\Sigma)$, $\varphi \vee \psi \in X$ iff either $\varphi \in X$ or $\psi \in X$.
(iv) For all $P\varphi \in \neg Cl(\Sigma)$, $P\varphi \in X$ iff either $\blacklozenge\varphi \in X$ or $\blacklozenge P\varphi \in X$.

Proof Claims (i)–(iii) are relatively straightforward, so we present the proof of claim (iv) and assume that $\mathsf{P}\varphi \in \neg Cl(\Sigma)$. For the forwards direction, assume that $\mathsf{P}\varphi \in X$. By the condition on the TLAE closure of a set, $\blacklozenge\mathsf{P}\varphi, \blacklozenge\varphi \in \neg Cl(\Sigma)$. Since $\mathsf{H}\varphi \leftrightarrow (\blacksquare\varphi \wedge \blacksquare\mathsf{H}\varphi)$ is an instance of axiom A9, it follows that $\vdash \mathsf{P}\varphi \leftrightarrow (\blacklozenge\mathsf{P}\varphi \vee \blacklozenge\varphi)$. Then, since $\mathsf{P}\varphi \in X$, one can infer that $X \vdash \blacklozenge\varphi \vee \blacklozenge\mathsf{P}\varphi$. Suppose that neither $\blacklozenge\varphi$ nor $\blacklozenge\mathsf{P}\varphi$ are in X, then, due to the definition of $\neg Cl(\Sigma)$ and claim (i), $\neg\blacklozenge\varphi, \neg\blacklozenge\mathsf{P}\varphi \in X$ and one can infer $X \vdash \neg(\blacklozenge\varphi \vee \blacklozenge\mathsf{P}\varphi)$, whence $X \vdash \bot$, which contradicts the fact that X is an atomic set over Σ.

For the other direction, assume that either $\blacklozenge\mathsf{P}\varphi \in X$ or $\blacklozenge\varphi \in X$. It follows from axiom A9 that $\vdash (\blacklozenge\mathsf{P}\varphi \vee \blacklozenge\varphi) \rightarrow \mathsf{P}\varphi$, which further implies that $X \vdash \mathsf{P}\varphi$, regardless of which case holds. By claim (ii), the assumption that $\mathsf{P}\varphi \in \neg Cl(\Sigma)$, and the assumption that $X \in At(\Sigma)$, we have that $\mathsf{P}\varphi \in X$. □

Lemma 2 *If $\varphi \in \neg Cl(\Sigma)$ and φ is consistent, then there is an $X \in At(\Sigma)$ such that $\varphi \in X$.*

Proof Similar to [9, Lemma 4.83]. □

We note that given a finite set X of formulae, we define $\widehat{X}$ to be a conjunction of all of its elements. Since all such conjunctions are equivalent according to our axiomatisation and semantics, we are free to use any conjunction of the elements of X for $\widehat{X}$.

Definition 11 (Canonical Model over Σ) Let Σ be a finite set of formulae. The *canonical model over* Σ is defined to be the tuple

$$\mathfrak{M}(\Sigma) := \langle W, \{W_{\mathfrak{d}^{\alpha_i}_j} \mid \mathfrak{d}^{\alpha_i}_j \in \mathtt{Wit}\}, \{W_{\mathfrak{e}^{\alpha_i}} \mid \alpha_i \in \mathtt{Agent}\}, R_{\Box}, R_{[\mathsf{A}]}, R_{\blacksquare}, R_{\mathsf{H}}\rangle$$

such that:

- $W := At(\Sigma)$
- $Y \in W_{\mathfrak{d}^{\alpha_i}_j}$ *iff* $\mathfrak{d}^{\alpha_i}_j \in Y$
- $Y \in W_{\mathfrak{e}^{\alpha_i}}$ *iff* $\mathfrak{e}^{\alpha_i} \in Y$
- $Y R_{\Box} Z$ *iff* $\widehat{Y} \wedge \Diamond\widehat{Z}$ is consistent
- $Y R_{[\mathsf{A}]} Z$ *iff* $\widehat{Y} \wedge \langle\mathsf{A}\rangle\widehat{Z}$ is consistent
- $Y R_{\blacksquare} Z$ *iff* $\widehat{Y} \wedge \blacklozenge\widehat{Z}$ is consistent
- $Y R_{\mathsf{H}} Z$ *iff* $Y R^{+}_{\blacksquare} Z$
- $V(p) := \{Y \in At(\Sigma) \mid p \in Y\}$

We take $R^{+}_{\blacksquare}$ to be the transitive closure of $R_{\blacksquare}$, and the definition of sets of moments associated with complex action types is as usual:

- $Y \in W_{t(\overline{\Delta}^{\alpha_i})}$ *iff* $Y \notin W_{t(\Delta^{\alpha_i})}$
- $Y \in W_{t(\Delta^{\alpha_i} \cup \Gamma^{\alpha_i})}$ *iff* $Y \in W_{t(\Delta^{\alpha_i})} \cup W_{t(\Gamma^{\alpha_i})}$.

Lemma 3 *Let* Σ *be a finite set of formulae and* $X \in At(\Sigma)$.

(i) If $[?] \in \{\Box, \blacksquare, [\mathsf{A}]\}$ *and* $\langle ? \rangle \in \{\Diamond, \blacklozenge, \langle \mathsf{A} \rangle\}$, *then for all* $\langle ? \rangle\varphi \in \neg Cl(\Sigma)$, $\langle ? \rangle\varphi \in X$ *iff there exists a* $Y \in At(\Sigma)$ *such that* $XR_{[?]}Y$ *and* $\varphi \in Y$.
(ii) If $P\varphi \in \neg Cl(\Sigma)$, *then* $P\varphi \in X$ *iff there exists a* Y *such that* $XR_H Y$ *and* $\varphi \in Y$.

Proof The two claims are proven similar to Lemma 4.86 and Lemma 4.89 in [9], respectively. □

Lemma 4 *Let* Σ *be a finite set of formulae. Then, the canonical model* $\mathfrak{M}$ *over* Σ *is a* TLAE*-model.*

Proof We know that $\mathfrak{M}(\Sigma)$ is an $\mathcal{L}_{\mathsf{TLAE}}$-model by definition. To prove the claim, it suffices to argue that $\mathfrak{M}(\Sigma)$ satisfies the properties of a TLAE-model. We only show the p(A3) and p(A6) cases as the remaining cases are similar or routine.

p(A3) Assume that $XR_{[\mathsf{A}]}Y$ and $XR_{[\mathsf{A}]}Z$ hold. We want to show that $Y = Z$, that is, we want to show that $\widehat{Y} \wedge \widehat{Z}$ is consistent (note that since Y and Z are atoms, $\widehat{Y}$ and $\widehat{Z}$ are jointly consistent *iff* $Y = Z$). We therefore assume that $\widehat{Y} \wedge \widehat{Z}$ is inconsistent and derive a contradiction. If $\widehat{Y} \wedge \widehat{Z}$ is inconsistent, then it follows that $\vdash \widehat{Y} \wedge \widehat{Z} \rightarrow \bot$. By modal reasoning, this implies that $\vdash \langle \mathsf{A} \rangle(\widehat{Y} \wedge \widehat{Z}) \rightarrow \langle \mathsf{A} \rangle\bot$. Observe that $\vdash \langle \mathsf{A} \rangle\widehat{Y} \wedge \langle \mathsf{A} \rangle\widehat{Z} \rightarrow \langle \mathsf{A} \rangle(\widehat{Y} \wedge \widehat{Z})$ holds, as it is a consequence of the axiom $\langle \mathsf{A} \rangle\varphi \rightarrow [\mathsf{A}]\varphi$; hence, $\vdash \langle \mathsf{A} \rangle\widehat{Y} \wedge \langle \mathsf{A} \rangle\widehat{Z} \rightarrow \langle \mathsf{A} \rangle\bot$. By modal and propositional reasoning, we have that $\vdash (\widehat{X} \wedge \langle \mathsf{A} \rangle\widehat{Y}) \wedge (\widehat{X} \wedge \langle \mathsf{A} \rangle\widehat{Z}) \rightarrow \bot$, meaning that $\vdash (\widehat{X} \wedge \langle \mathsf{A} \rangle\widehat{Y}) \rightarrow \bot \vee (\widehat{X} \wedge \langle \mathsf{A} \rangle\widehat{Z}) \rightarrow \bot$. The last theorem implies that either $XR_{[\mathsf{A}]}Y$ or $XR_{[\mathsf{A}]}Z$ does not hold (by the definition or $R_{[\mathsf{A}]}$), thus contradicting our assumption.

p(A6) Let $X \in W, \alpha_i$ an agent, and suppose that there exists a Y such that $XR_{\Box}Y$ and $Y \in W_{\mathfrak{e}^{\alpha_i}}$. We want to show that there exists a $Z \in W$ such that $XR_{\Box}Z$ and $Z \notin W_{\mathfrak{e}^{\alpha_i}}$. By Definition 8, we know that $\Diamond\neg\mathfrak{e}^{\alpha_i} \in \neg Cl(\Sigma)$. We aim to show that $\widehat{X} \wedge \Diamond\neg\mathfrak{e}^{\alpha_i}$ is consistent, since this will imply that $\Diamond\neg\mathfrak{e}^{\alpha_i} \in X$, due to the fact that X is an atomic set. Thus, we assume that $\vdash \widehat{X} \wedge \Diamond\neg\mathfrak{e}^{\alpha_i} \rightarrow \bot$ to derive a contradiction. By axiom A6, we have $\vdash \widehat{X} \wedge \Diamond\mathfrak{e}^{\alpha_i} \rightarrow \bot$ as a consequence, which implies $\vdash \widehat{X} \wedge \Diamond\mathfrak{e}^{\alpha_i} \wedge \Diamond\widehat{Y} \rightarrow \bot$ by propositional reasoning. Modal and propositional reasoning may then be applied to derive the following

$$\vdash \widehat{X} \wedge \Diamond(\mathfrak{e}^{\alpha_i} \wedge \widehat{Y}) \rightarrow \bot$$

We know that $\widehat{Y} \wedge \mathfrak{e}^{\alpha_i}$ is consistent because $Y \in W_{\mathfrak{e}^{\alpha_i}}$. However, since Y is an atomic set and $\mathfrak{e}^{\alpha_i} \in \neg Cl(\Sigma)$, it follows that $\widehat{Y} \wedge \mathfrak{e}^{\alpha_i}$ is equivalent to $\widehat{Y}$. Hence,

$$\vdash \widehat{X} \wedge \Diamond\widehat{Y} \rightarrow \bot$$

contradicting our assumption that $XR_{\Box}Y$. Consequently, $\Diamond\neg\mathfrak{e}^{\alpha_i} \in X$, so by Lemma 3, there exists a $Z \in At(\Sigma)$ such that $XR_{\Box}Z$ and $\neg\mathfrak{e}^{\alpha_i} \in Z$. The latter fact implies that $Z \notin W_{\mathfrak{e}^{\alpha_i}}$ by Definition 11. □

Lemma 5 *(Truth Lemma) Let $\mathfrak{M}(\Sigma)$ be the canonical model over Σ. For all atomic sets Y and all $\varphi \in \neg Cl(\Sigma)$, $\mathfrak{M}(\Sigma), Y \models \varphi$ iff $\varphi \in Y$.*

Proof By induction on the complexity of φ. □

Theorem 4 *(Weak Completeness) For any formula φ, if $\models \varphi$, then $\vdash \varphi$.*

Proof Follows from Lemma 4 and Lemma 5. □

5 Formal Notions of Instrumentality

In this section, we formalise a variety of instrumentality notions corresponding to the philosophical analysis of Section 2. As will be demonstrated, the logic TLAE suffices to capture many of the desired nuances present in judgments of instrumentality. In Section 5.1, we show how time intervals can be utilised to evaluate an action's suitability for serving a particular purpose (criterion I). In Section 5.2, using the various definitions of instrumentality, we semantically define a collection of value judgments qualifying instruments as 'better', 'best', 'worst', 'good', and 'poor' (criterion II). In Section 5.3, we show that, as a result of using past and future references, the obtained formal definitions capture the inherent defeasible nature of judgments of instrumentality (criterion III).

5.1 Elementary Instrumentality Notions

In [7], four formal definitions of instrumentality are provided: a 'basic' and a 'proper' notion of instrumentality, which are either agent-dependent or agent-independent. (The terms 'basic' and 'proper' in [7] refer to our account of instruments in parts (1) and (2) of Definition 3, respectively, i.e., 'plain' and 'excellent' instruments.) As remarked above, the logical system LAE, introduced in [7], is closely related to the logic TLAE presented in this paper. The principal difference between LAE and TLAE is that the former's language does not include modalities that refer to the past, whereas the latter does. To be precise, the language of TLAE includes the additional ■ and H operators. These operators allow us to syntactically define the notion of '*candidate instrument*' and '*excellent candidate instrument*' (see Definition 1), referring either to a finite interval of time preceding the moment of evaluation or to the entire past of the moment of evaluation. These notions were merely semantically defined in [7]. The logical characterisation of the ■ operator enables us to count successful applications of candidate instruments and syntactically capture intervals of time.

In (d11) below, we formalise the agent-dependent notion of candidate φ-instrument as presented in item (1) of Definition 1. Observe that the definition is relativised to a past time interval with length n.

Candidate Instrument for α (with a past interval of length n)

$$[\Delta^\alpha]_n^{c-instr}\varphi := \bigvee_{0\leq i\leq n} \blacklozenge^i(t(\Delta^\alpha)\wedge \blacklozenge[\Delta^\alpha]^{would}\varphi) \tag{d11}$$

The formula $[\Delta^{\alpha}]_{n}^{c-instr}\varphi$ (d13) reads as 'somewhere within the past interval of length n there is a moment v at a distance of i units of time that witnessed the successful performance of Δ by agent α such that at v's immediate predecessor, the performance of Δ by that agent would have guaranteed φ'.

Agent-dependent excellent candidate φ-instruments are, then, formalised by combining the above definition with the idea that *every* past performance of the relevant action type has led to the intended outcome (item (2) of Definition 1).

Excellent Candidate Instrument for α (*with a past interval of length* n)

$$[\Delta^{\alpha}]_{n}^{ex.c-instr}\varphi := [\Delta^{\alpha}]_{n}^{c-instr}\varphi \wedge \bigwedge_{1\leq k\leq n} \blacksquare^{k}[\Delta^{\alpha}]^{would}\varphi \qquad \text{(d12)}$$

We read (d12) as 'action type Δ has proven to be a candidate instrument for φ for α at least once in the interval, and every other performance of Δ by α within the interval would have also guaranteed φ'. One can say that, within the past interval n, the action type Δ has a one hundred percent success rate for α in obtaining φ. In the sequel, we introduce ways of refining these definitions.[11]

Agent-independent generalizations of (d11) and (d12) are captured by (d13), respectively (d14) (cf. the agent-dependent definitions of [7]).

$$[\Delta]_{n}^{c-instr}\varphi := \bigvee_{\alpha\in\texttt{Agent}}\ \bigvee_{0\leq i\leq n} \blacklozenge^{i}(t(\Delta^{\alpha}) \wedge \blacklozenge[\Delta^{\alpha}]^{would}\varphi) \qquad \text{(d13)}$$

$$[\Delta]_{n}^{ex.c-instr}\varphi := [\Delta]_{n}^{c-instr}\varphi \wedge \bigwedge_{1\leq k\leq n} \blacksquare^{k} \bigwedge_{\alpha\in\texttt{Agent}} [\Delta^{\alpha}]^{would}\varphi \qquad \text{(d14)}$$

Henceforth, we focus on *agent-dependent* notions. The agent-independent versions of each of the definitions below can be straightforwardly obtained.

The more refined definitions of instruments and excellent instruments—corresponding to items (1) and (2) of Definition 3—are obtained by adding the pivotal conjectural element, reflecting the agent's expectations about the instrument's suitability in the immediate future. To capture these notions, we first need to alter the agentive operator *would* (d1) (Section 3.2).

Expected Would

$$[\Delta^{\alpha}]_{ex}^{would}\varphi := \Box((t(\Delta^{\alpha}) \wedge \mathfrak{e}^{\alpha}) \rightarrow \varphi). \qquad \text{(d15)}$$

The 'expected would' operator (d15) restricts the formula's evaluation to immediate future moments that the agent *expects* as continuations of the present. Using (d15),

[11] Observe that in (d11) and (d12) reference to past experience—i.e., $\blacklozenge^{i}$ and $\blacksquare^{i}$—also includes reference to outcomes obtained at the moment of evaluation w. Otherwise, a performance of Δ immediately before w might fail to deliver φ at w.

we formalise items (1) and (2) of Definition 3 as follows:

$$\textit{Instrument for } \alpha \textit{ (with a past interval of length } n)$$

$$[\Delta^{\alpha}]_n^{instr}\varphi := \bigvee_{0\leq i\leq n} \blacklozenge^i (t(\Delta^{\alpha}) \wedge \blacklozenge[\Delta^{\alpha}]^{would}\varphi) \wedge [\Delta^{\alpha}]_{ex}^{would}\varphi \tag{d16}$$

$$\textit{Excellent Instrument for } \alpha \textit{ (with a past interval of length } n)$$

$$[\Delta^{\alpha}]_n^{ex-instr}\varphi := [\Delta^{\alpha}]_n^{instr}\varphi \wedge \bigwedge_{1\leq k\leq n} \blacksquare^k[\Delta^{\alpha}]^{would}\varphi \tag{d17}$$

The final formalisations (d16) and (d17)—to which we sometimes refer as 'proper instruments'—differ from their candidate counterparts (d11) and (d12) through the additional conjunct expressing that the agent expects that, at the moment of evaluation, she would guarantee φ by performing Δ. Stronger notions of (excellent) instruments are straightforwardly obtained by using a definition of 'expected could' instead of 'expected would' as the last conjunct in (d16).[12] The following implications are theorems of TLAE and show the various relations between the four notions defined so far.

$$[\Delta^{\alpha}]_n^{ex-instr}\varphi \rightarrow [\Delta^{\alpha}]_n^{ex.c-instr}\varphi \;\; \textit{and} \;\; [\Delta^{\alpha}]_n^{instr}\varphi \rightarrow [\Delta^{\alpha}]_n^{c-instr}\varphi$$

$$[\Delta^{\alpha}]_n^{ex.c-instr}\varphi \rightarrow [\Delta^{\alpha}]_n^{c-instr}\varphi \;\; \textit{and} \;\; [\Delta^{\alpha}]_n^{ex-instr}\varphi \rightarrow [\Delta^{\alpha}]_n^{instr}\varphi$$

Before moving to comparative notions of instrumentality, we make three remarks. First, purposes may include action formulae or action witnesses. For instance, agent α's purpose may be to ensure that agent β could bring about φ by performing Δ. In that case, α is looking for the action that will ensure that, at the next moment, $[\Delta^{\beta}]\varphi$ holds. We will not further pursue this here.

Second, we did not consider deliberative versions of instrumentality. Definitions (d11), (d12), (d16), and (d17) will qualify each action type as an instrument for bringing about *tautologous* propositions. Deliberative variants can be straightforwardly obtained in the spirit of [5]'s deliberative STIT operator, e.g., if φ is the purpose at hand, the possibility of $\neg\varphi$ is required.

Third, the volitional concepts of Section 3.4 (e.g., 'producing' and 'destroying') can also be employed in the context of instrumentality. Such definitions can be straightforwardly given in the framework of TLAE. For instance, (d18) (below) expresses that, for agent α, Δ is an instrument for *producing* p because (i) α produced p through performing Δ at least once in the past (with an interval of length n) and (ii) α *expects*

[12] Such stronger notions are used in [7]. As pointed out by a reviewer, the use of expected could can be problematic. For instance, suppose that at w agent α acknowledges a relation between past performances of action Δ (e.g., turning a car's ignition key) and an outcome φ (the car's motor is running). Now, suppose that at w, α expects that Δ will serve purpose φ in the immediate future. This may suffice for α to consider Δ an instrument for φ at w, even if Δ cannot be performed at w for reasons unknown to α (e.g., the car's battery is down).

to produce p through Δ at present (the four action types require atoms in their formulation, see Section 2.1.).

$$[\Delta^{\alpha}]^{instr}_{prod-n} p := \begin{array}{ll} (i) & \bigvee_{0 \leq i \leq n} \blacklozenge^{i}(t(\Delta^{\alpha}) \wedge \blacklozenge[\Delta^{\alpha}]^{would}_{prod} p) \wedge \\ (ii) & [\Delta^{\alpha}]^{would}_{ex} p \wedge \neg p \wedge \Diamond(\neg p \wedge \mathfrak{e}^{\alpha}) \end{array} \tag{d18}$$

That (d18) is a stronger notion than (d16) follows from the following theorem:

$$[\Delta^{\alpha}]^{instr}_{prod-n} p \rightarrow [\Delta^{\alpha}]^{instr}_{n} p$$

5.2 Good Instruments and Comparative Instrumentality

As shown in the previous section, our language $\mathcal{L}_{\mathsf{TLAE}}$ is suitable for the definition of candidate instruments, as per (d11) and (d12), as well as proper instruments, as per (d16) and (d17). We now discuss how to evaluate different instruments serving the same purpose, giving rise to notions such as 'better' and 'good' instruments (see Definition 2 and items (3) and (4) of Definition 3). In our framework, we can use TLAE-models to determine whether an available (candidate) instrument is regarded as better than another. In this section, we proceed with a semantic analysis of comparative instrumentality, providing various notions of *better (worse), best (worst),* and *good (poor)* instruments.

We adopt the following methodology:

1. Collect all the instruments that serve φ;
2. Determine the success ratio of the obtained instruments w.r.t. φ.

For comparative judgments of *betterness*, we proceed accordingly:

3. Order the available instruments on the basis of their success ratio;
4. Identify the best and the worst;
5. Identify better instruments by comparing instruments within their order.

For comparative judgments of instrumental *goodness*, we proceed as follows:

6. Identify good instruments by checking whether their success ratio satisfies a certain threshold ratio;
7. Identify good instruments by checking whether they satisfy certain additional thresholds, such as a minimum amount of past witnesses.

We addressed step (1) in the previous section via (d16). In Section 5.2.1, we deal with (3)-(5), and in Section 5.2.2, we deal with (6) and (7). First, let us address the notion of *success ratio* from step (2). For readability, in what follows, we omit explicit reference to a past interval of length n without a loss of generality. Recall from page 22 that TLAE-frames are rooted tree-like structures, meaning each moment w has a finite past. Thus, there are only finitely many witnesses of an instrument's application. We use '$*$' to denote consideration of the *entire past*. That is, for a moment $w \in W$, $w \models [\Delta^{\alpha}]^{instr}_{*} \varphi$ signifies that we evaluate the past of w up to the root of the model.

In the sequel, we provide our definitions for proper instruments only, that is, using (d16) (variations can be straightforwardly obtained). Let φ be the purpose endorsed by agent α at moment w. Then, we let $Instr^{w}_{\alpha}(\varphi)$ denote the set containing all φ-instruments available to α at w with respect to the entire past of w, i.e., all actions that served φ at least once, and of which α expects that they will serve φ again.[13] Definition (d19) below makes this formally precise.

$$\begin{aligned}&Available\ \varphi - instruments\ \ for\ \alpha \in \texttt{Agent}\ \ at\ w\\ &Instr^{w}_{\alpha}(\varphi){:=}\{\Delta \in \mathcal{L}_{\mathsf{Act}} \mid w \models [\Delta^{\alpha}]^{instr}_{*}\varphi\}\end{aligned} \tag{d19}$$

For each collected available instrument $\Delta \in Instr^{w}_{\alpha}(\varphi)$, we must determine how 'successful' it is at securing φ. Here, we opt for defining success in terms of an achievement/failure ratio. Below, *achievement* (d20) refers to the fact that by performing Δ agent α would secure φ and, in fact, did secure φ.

$$\begin{aligned}&Achievement\ \ of\,\Delta \in Instr^{w}_{\alpha}(\varphi)\\ &Achiev^{w}_{\alpha}(\varphi,\Delta){:=}\{\theta \mid \theta = \blacklozenge^{i}(t(\Delta^{\alpha}) \wedge \blacklozenge[\Delta^{\alpha}]^{would}\varphi), 0{\leq}i, and w \models \theta\}\end{aligned} \tag{d20}$$

In (d20), we use $0 \leq i$ to indicate that the moment of evaluation may serve as a witness of a successful application of Δ initiated in its immediate preceding moment. Furthermore, the first conjunct in θ witnesses the actual performance of Δ and attainment of φ, whereas the second conjunct ensures that the simultaneous occurrence of Δ and φ is not merely coincidental.

We define *failure* (d21) in terms of the expectations of the agent involved. A failure to secure φ by α's performance of Δ at w means that (i) at w, Δ has just been performed by α, (ii) φ does not hold at w, and (iii) at the immediately preceding moment α expected that by performing Δ, φ would be attained.

$$\begin{aligned}&Failure\ \ of\,\Delta \in Instr^{w}_{\alpha}(\varphi)\\ &Fail^{w}_{\alpha}(\varphi,\Delta){:=}\{\theta \mid\theta = \blacklozenge^{i}(t(\Delta^{\alpha}) \wedge\neg\varphi \wedge \blacklozenge[\Delta^{\alpha}]^{would}_{ex}\varphi), 0{\leq}i,\ \ and\ \ w \models \theta\}\end{aligned} \tag{d21}$$

Through the use of $\blacklozenge^{i}$, all members of $Achiev^{w}_{\alpha}(\varphi,\Delta)$ and $Fail^{w}_{\alpha}(\varphi,\Delta)$ represent unique past moments. In what follows, we sometimes abuse notation and let these sets stand for the sets of moments witnessing the respective achievements and failures. Furthermore, the two sets are finite since the past is finite, and so, the cardinality of the two sets in (d20) and (d21) can be utilised to count the amount of achieved or failed applications of the instrument under consideration. Thus, we have the means

[13] Note that even if $Instr^{w}_{\alpha}(\varphi)$ contains infinitely many action types, up to logical equivalence, the set will only contain finitely many. We come back to this in Section 5.2.1.

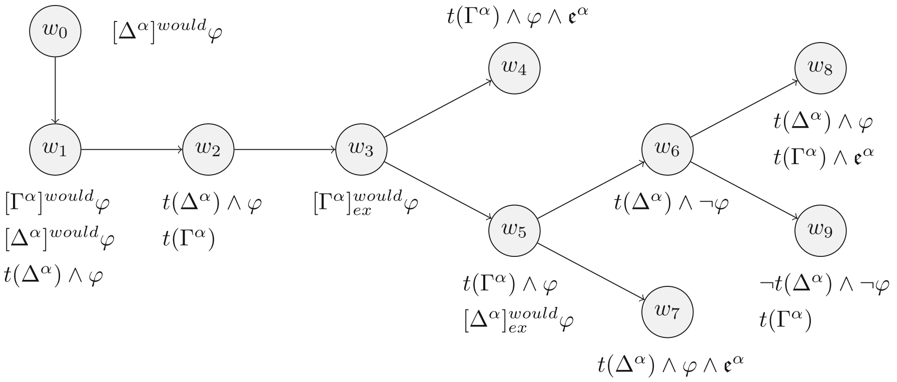
w_0
$[\Delta^\alpha]^{would}\varphi$
$t(\Gamma^\alpha) \wedge \varphi \wedge \mathfrak{e}^\alpha$
w_4
w_8
w_1
w_2
w_3
w_6
$t(\Delta^\alpha) \wedge \varphi$
$t(\Gamma^\alpha) \wedge \mathfrak{e}^\alpha$
$[\Gamma^\alpha]^{would}\varphi$
$[\Delta^\alpha]^{would}\varphi$
$t(\Delta^\alpha) \wedge \varphi$
$t(\Delta^\alpha) \wedge \varphi$
$t(\Gamma^\alpha)$
$[\Gamma^\alpha]^{would}_{ex}\varphi$
$t(\Delta^\alpha) \wedge \neg\varphi$
w_5
w_9
$t(\Gamma^\alpha) \wedge \varphi$
$[\Delta^\alpha]^{would}_{ex}\varphi$
w_7
$\neg t(\Delta^\alpha) \wedge \neg\varphi$
$t(\Gamma^\alpha)$
$t(\Delta^\alpha) \wedge \varphi \wedge \mathfrak{e}^\alpha$

Fig. 4 An example of evaluating φ-instruments Γ and Δ for α at w_6

to calculate a *success ratio* for the collected instruments, i.e., (d22). We denote the cardinality of a set S by $|S|$.

$$\begin{aligned} &\textit{Success ratio of } \Delta \in \textit{Instr}^w_\alpha(\varphi) \\ &\textit{Succ}^w_\alpha(\varphi, \Delta) := \frac{|\textit{Achiev}^w_\alpha(\varphi, \Delta)|}{|\textit{Achiev}^w_\alpha(\varphi, \Delta) \cup \textit{Fail}^w_\alpha(\varphi, \Delta)|} \end{aligned} \tag{d22}$$

To exemplify the above machinery, consider Fig. 4 and suppose there is an agent α at moment w_6 who aims to bring about φ. There are two available instruments at w_6, namely, Δ and Γ, whose potential was witnessed at w_2. We find that α expects both instruments to serve φ at w_6 (see w_8), although α's current expectations about Γ are incorrect (see w_9). Furthermore, notice that, from the vantage point of w_6, Γ and Δ both have two achievement witnesses, namely, $\textit{Achiev}^{w_6}_\alpha(\varphi, \Gamma) = \{w_2, w_5\}$, and $\textit{Achiev}^{w_6}_\alpha(\varphi, \Delta) = \{w_1, w_2\}$. Furthermore, Δ also has a witness of a failed application at w_6, that is, $\textit{Fail}^{w_6}_\alpha(\varphi, \Delta) = \{w_6\}$. At w_5, α had expectations that turned out to be wrong at w_6. Hence, the success ratios of Γ and Δ at w_6 are $\textit{Succ}^{w_6}_\alpha(\varphi, \Gamma) = 1$ and $\textit{Succ}^{w_6}_\alpha(\varphi, \Delta) = \frac{2}{3}$, respectively.

5.2.1 Comparative Judgments of Betterness

The notion of success ratio is pivotal for defining various types of axiological—i.e., value—judgment concerning instrumentality. In particular, we are interested in the following axiological concepts: *best*, *worst*, *better*, *good*, and *poor*. The latter two are formally addressed in the next section.

Recall that the language $\mathcal{L}_{\mathsf{TLAE}}$ allows only for finitely many action types, up to provable equivalence of the formulas that witness their performance by an agent. That is, since $\mathcal{L}_{\mathsf{Act}}$ is constructed over a finite number of atomic action types, for each agent there will be finitely many equivalence classes $[\![\Delta^\alpha]\!] := \{\Gamma^\alpha \mid \vdash_{\mathsf{TLAE}} t(\Delta^\alpha) \leftrightarrow t(\Gamma^\alpha)\}$ of equivalent actions. We let $\mathsf{EqAct} = \{[\![\Delta^\alpha]\!] \mid \Delta \in \mathcal{L}_{\mathsf{Act}} \text{ and } \alpha \in \texttt{Agent}\}$ be the set of all such equivalence classes. Consequently, we obtain a finite ordering of actions (up to equivalence) when ordering candidate instruments. This observation enables us to identify those instruments at the ordering's upper- and lower-bound.

Naive best, worst, and better. We define $\succeq_\varphi^w$ to be a success ratio ordering over actions such that, for each $\Delta, \Gamma \in \mathit{Instr}_\alpha^w(\varphi)$, we have

$$\Delta^\alpha \succeq_\varphi^w \Gamma^\alpha \textit{ iff } \mathit{Succ}_\alpha^w(\varphi, \Delta) \geq \mathit{Succ}_\alpha^w(\varphi, \Gamma). \tag{d23}$$

We define $\Delta^\alpha \succ_\varphi^w \Gamma^\alpha$ as the conjunction $\Delta^\alpha \succeq_\varphi^w \Gamma^\alpha$ and $\Gamma^\alpha \not\succeq_\varphi^w \Delta^\alpha$. We interpret $\succeq_\varphi^w$ as a "betterness" relation: i.e., we read $\Delta^\alpha \succeq_\varphi^w \Gamma^\alpha$ as 'at w, Δ is a *weakly better* instrument for agent α to secure φ than the instrument Γ' (i.e., Δ is at least as good as Γ for α in order to get φ). Then, $\Delta^\alpha \succ_\varphi^w \Gamma^\alpha$ expresses that 'Δ is a (strictly) *better* φ-instrument than Γ, for α at w'.

Likewise, we can define notions of best and worst because, in the ordering, we find an upper and lower bound. Since we have a finite ordering of classes of equivalent actions, we know that if $\mathit{Instr}_\alpha^w(\varphi) \neq \emptyset$, there are $\Delta, \Gamma \in \mathit{Instr}_\alpha^w(\varphi)$ such that Δ^α is an upper bound, and Γ^α is a lower bound of $\succeq_\varphi^w$. In other words, there are no $\Theta, \Sigma \in \mathit{Instr}_\alpha^w(\varphi)$ such that $\Theta^\alpha \succ_\alpha^w \Delta^\alpha$, respectively $\Gamma^\alpha \succ_\alpha^w \Sigma^\alpha$. Note that it can be that $\mathit{Succ}_\alpha^w(\varphi, \Delta) = \mathit{Succ}_\alpha^w(\varphi, \Gamma)$ or that $\Delta = \Gamma$. Naively, we may say that an instrument $\Delta \in \mathit{Instr}_\alpha^w(\varphi)$ in an upper bound of $\succeq_\varphi^w$ is *among the best* instruments for α at w, whereas an instrument $\Gamma \in \mathit{Instr}_\alpha^w(\varphi)$ in a lower bound of $\succeq_\varphi^w$ is *among the worst* instruments for α at w for securing φ.

There is an obvious objection to this naive approach: Suppose there are only two available φ-instruments $\Delta, \Gamma \in \mathit{Instr}_\alpha^w(\varphi)$ for α at w such that $\mathit{Succ}_\alpha^w(\varphi, \Delta) = 1$ and $\mathit{Succ}_\alpha^w(\varphi, \Gamma) = 0.999$. Our naive definition tells us: (i) Δ is better than Γ (for α, in securing φ at w), (ii) Δ is among the best available φ-instruments, and (iii) Γ is among the worst available φ-instruments. That Γ is the 'worst' instrument is arguably an overstatement. In fact, the argument can even be made that Γ is an exceptionally better φ-instrument than Δ. For example, suppose that Δ was performed only once in the past by α, securing φ and that Γ was performed 1000 times, securing φ 999 times. Then, one should in fact be able to conclude that Γ is the best instrument. In particular, Γ has proven to be more *reliable* in producing the desired outcome. This observation motivates the instalment of *thresholds*, which filter out insufficient experience, ensuring a certain quality standard of the instruments evaluated.

Thresholds for best, worst, and better. We consider two types of thresholds. First, we can impose a threshold n that states the minimum amount of past witnesses of an instrument's application. Second, we can impose a threshold on the minimum success ratio of potential instruments. We begin by considering the first threshold and adopt the second approach in Section 5.2.2.

Let n refer to the threshold that needs to be met by the total amount of witnesses $|\mathit{Achiev}_\alpha^w(\varphi, \Delta) \cup \mathit{Fail}_\alpha^w(\varphi, \Delta)|$. We write $\mathit{Instr}_\alpha^w(\varphi, n)$ to denote the set of φ-instruments satisfying threshold n, as defined in (d24).

$$\begin{aligned}&\textit{Available } \varphi\textit{-instruments for } \alpha \in \texttt{Agent} \textit{ at } w \textit{ (with threshold } n \in \mathbb{N})\\&\mathit{Instr}_\alpha^w(\varphi, n) := \{\Delta \mid \Delta \in \mathit{Instr}_\alpha^w(\varphi) \textit{ and } |\mathit{Achiev}_\alpha^w(\varphi, \Delta) \cup \mathit{Fail}_\alpha^w(\varphi, \Delta)| \geq n\}\end{aligned} \tag{d24}$$

Based on (d24), we can refine the notion of ‘betterness’ as follows:

$$\begin{aligned}&\textit{Better } \varphi\textit{–instruments, relative to threshold } n \in \mathbb{N}\\&\textit{For each}\Delta, \Gamma \in \mathit{Instr}^w_\alpha(\varphi, n),\ \Delta^\alpha \succeq^w_{\varphi,n} \Gamma^\alpha \textit{ iff } \mathit{Succ}^w_\alpha(\varphi, \Delta) \geq \mathit{Succ}^w_\alpha(\varphi, \Gamma)\end{aligned} \tag{d25}$$

We interpret $\Delta^\alpha \succeq^w_{\varphi,n} \Gamma^\alpha$ in (d25) as ‘Δ^α is a *weakly better* φ-instrument than Γ^α at w given threshold n’. Imposing a threshold installs a quality control in providing axiological judgments of instrumentality. The undesirable consequences of the 0.999 success rate example in the previous section can now be excluded by stipulating a threshold of any value $n > 1$, identifying Γ as the best φ-instrument for α. In (d26) and (d27) below, we define the sets $\mathit{Best}^w_\alpha(\varphi, n)$ for ‘among the best’ and $\mathit{Worst}^w_\alpha(\varphi, n)$ for ‘among the worst’, which contain those instruments in the upper, respectively lower bound of the ordering, satisfying the imposed threshold.

$$\begin{aligned}&\textit{Best } \varphi - \textit{instruments, relative to threshold } n \in \mathbb{N}\\&\mathit{Best}^w_\alpha(\varphi, n){:=}\{\Delta \in \mathit{Instr}^w_\alpha(\varphi, n) \mid \ \textit{for each } \Gamma \in \mathit{Instr}^w_\alpha(\varphi, n),\ \Delta^\alpha \succeq^{w,n}_\varphi \Gamma^\alpha\}\end{aligned} \tag{d26}$$

$$\begin{aligned}&\textit{Worst } \varphi\textit{–instruments, relative to threshold } n \in \mathbb{N}\\&\mathit{Worst}^w_\alpha(\varphi, n){=}\{\Delta \in \mathit{Instr}^w_\alpha(\varphi, n) \mid \textit{for each } \Gamma \in \mathit{Instr}^w_\alpha(\varphi, n),\ \Gamma^\alpha \succeq^{w,n}_\varphi \Delta^\alpha\}\end{aligned} \tag{d27}$$

Reconsider the example in Fig. 4. Depending on the threshold applied, either Δ or Γ will qualify as among the best (worst) φ-instruments at w_6. For instance, a threshold of $n = 3$ excludes Γ as a potential ‘best’ φ-instrument, i.e., $\Gamma \in \mathit{Instr}^{w6}_\alpha(\varphi, 2)$, but $\Gamma \notin \mathit{Instr}^{w6}_\alpha(\varphi, 3)$. In fact, we find that $\Gamma \in \mathit{Best}^{w6}_\alpha(\varphi, 2)$ and $\Delta \in \mathit{Worst}^{w6}_\alpha(\varphi, 2)$, whereas $\Delta \in \mathit{Best}^{w6}_\alpha(\varphi, 3)$. Furthermore, observe that $\Delta \notin \mathit{Worst}^{w6}_\alpha(\varphi, 3)$ since for $\Theta = \Delta \cup \Gamma$ we have $\Theta \in \mathit{Instr}^{w6}_\alpha(\varphi, 3)$ and $\mathit{Succ}^{w6}_\alpha(\varphi, \Delta) > \mathit{Succ}^{w6}_\alpha(\varphi, \Theta)$.

5.2.2 Thresholds and Good Instruments

In this section, we suggest possible ways to formally address the assessment of *good*-instruments (item (4) of Definition 3). As discussed in Section 2, for von Wright, comparative judgments are *objective* since they depend on empirical data (e.g., collecting past experiences) and logical orderings (e.g., success ratio orderings). However, such judgments become more problematic when we address the label ‘good’.

Von Wright determines good instruments on the basis of their ‘good-making properties’. Recall, how *well* a knife cuts depends on the good-making property associated with ‘cutting’: the sharpness of the knife. Such value judgments concerning knives, von Wright argues, can be objectively assessed, especially when we order an available set of knives according to their sharpness. Without such a comparative set, things become more difficult since we need to define a minimal degree of ‘sharpness’ that enables ‘cutting well’. Von Wright addresses this degree through *causal properties*

[38, p. 26], which determine the causal relation between 'sharpness' and 'cutting'—a relationship that may be objectively assessed through empirical investigations. Then, a knife cuts 'well' or qualifies as a *good* cutting-instrument, whenever it has the causal property of sharpness needed for cutting (note that this property depends on what needs to be cut).

We mention two problems related to the above approach. First, such judgments of *goodness* are a special case of comparative judgments. That is, the goodness reflected in a causal property is a goodness relative to (compared to) a threshold, for instance, the minimum amount of sharpness for cutting vegetables. In other words, judgments of instrumental goodness remain comparative but objective with respect to an external criterion, i.e., a threshold. Therefore, the term 'good' employed here does not differ from any other use of 'good' which, for instance, is defined relative to some theory of ethics. Still, the fact that such a threshold (the causal property) is rooted in experience allows for the judgment to be objectively evaluated. This observation is compatible with von Wright's ideas due to his emphasis on the logical nature of such judgments.

Second, for an agent in a time-limited decision-making situation, determining the causal properties of an instrument (such as sharpness) may be an overburdening task. An agent that aims at cutting, say, a piece of paper may not have access to determining the causal properties of a different knife's sharpness relative to the robustness of the piece of paper. Although causal properties form objective criteria for obtaining accurate judgments of instrumentality from a theoretical point of view, from the agent's point of view, such criteria are unrealistic and impractical. We find this to be a strong reason for using the agent's *personal experience* with an instrument (such as the knives available for cutting) as a criterion for judging instrumentality.

The question that remains is: What qualifies as sufficient experience for judging an instrument as *good* for a given purpose? We provide two notions of threshold that serve to denote sufficient experience: first, we propose a threshold on the success ratio of a given instrument and, second, we combine the former with the notion of a minimum amount of past witnesses (see Section 5.2.1).

$$Good_n\ \varphi{-}instruments,\ for\ \alpha \in \texttt{Agent}\ at\ w\ (with\ n \in \mathbb{R}\ and\ 0 \leq n \leq 1)$$

$$Good^w_\alpha(\varphi, n) := \{\Delta \mid \Delta \in Instr^w_\alpha(\varphi)\ and\ Succ^w_\alpha(\varphi, \Delta) \geq n\} \tag{d28}$$

Definition (d28) tells us that an action Δ is a $good_n$ φ-instrument for α at w, whenever 'Δ qualifies as a φ-instrument for α at w, and its success ratio satisfies threshold n'. Observe that the initial definition of (excellent) instruments (d16) and (d17)—i.e., items (1) and (2) of Definition 3—are limiting cases of $good_n$ instruments, namely, where the success ratios are $n > 0$ and $n = 1$, respectively. The relation between 'good', and (d16) and (d17) is expressed through the following:

$$\Delta \in Good^w_\alpha(\varphi, 1)\ iff\ w \models [\Delta^\alpha]^{ex-instr}_*\varphi\ iff\ \Delta \in Best^w_\alpha(\varphi, 1)$$

$$\Delta \in Good^w_\alpha(\varphi, n)\ \ for\ \ some\ \ 0 < n \leq 1\ iff\ w \models [\Delta^\alpha]^{instr}_*\varphi$$

A *poor* instrument can be defined in two ways: it is either an instrument failing to meet a 'poorness'-threshold n or an instrument that fails to qualify as good. We opt for the first approach—represented by (d29)—to leave room for instruments that are considered neither good nor poor.

$$\begin{aligned}&Poor_n\ \varphi\text{–}instruments,\ \ for\ \alpha \in \texttt{Agent}\ at\ w\ (n \in \mathbb{R}\ and\ 0 \leq n \leq 1)\\ &Poor^w_\alpha(\varphi, n){:=}\{\Delta \mid \Delta \in Instr^w_\alpha(\varphi)\ and\ Succ^w_\alpha(\varphi, \Delta) \leq n\}\end{aligned} \tag{d29}$$

We face the same objection encountered while discussing naive betterness in Section 5.2.1. Here too, an instrument that proved itself a certain number of times could be considered more *reliable* and thus more eligible for the label 'good'. Therefore, in some cases, we may need to expand definition (d28) with a second threshold imposing a minimum amount of experience with the instrument in question. The resulting definition is presented below (d30). We read $\Delta \in Good^w_\alpha(\varphi, n, m)$ as 'at w for α, action Δ is a *good* φ-instrument having a success ratio meeting n and a minimum amount of past witnesses m'.

$$\begin{aligned}&Good^m_n\ \varphi\text{–}instruments,\ \ for\ \alpha \in \texttt{Agent}\ \ at\ w\ \ (n \in \mathbb{R}, 0 \leq n \leq 1, m \in \mathbb{N})\\ &Good^w_\alpha(\varphi, n, m){:=}\{\Delta \mid \Delta \in Instr^w_\alpha(\varphi, m)\ and\ Succ^w_\alpha(\varphi, \Delta) \geq n\}\end{aligned} \tag{d30}$$

The modified definition of 'poor' instruments is similarly obtained from (d29).

Last, as an illustration of the above machinery, consider the TLAE-model provided in Fig. 4. Suppose agent α desires to secure φ at w_6. Then, suppose the minimal required success ratio is $n = 0.75$. In that case, at w_6 we find that only $\Gamma \in Good^{w_6}_\alpha(\varphi, n)$ since $Succ^{w_6}_\alpha(\varphi, \Gamma) = 1 \geq 0.75$. Action Δ fails to qualify due to $Succ^{w_6}_\alpha(\varphi, \Delta) = \frac{2}{3} < 0.75$. Suppose we impose the additional constraint that the minimum amount of witnesses is $3 = m$. In that case, both $\Gamma, \Delta \notin Good^{w_6}_\alpha(\varphi, n, m)$ since Γ does not satisfy m and Δ doesn't satisfy n. Still, despite Δ not being able to qualify as a 'good' φ-instrument, it is nevertheless the 'best' φ-instrument for α at w_6, i.e., $\Delta \in Best^{w_6}_\alpha(\varphi, m)$.

In conclusion, whether an instrument is considered 'good' or 'poor' depends on its evaluative criteria. Nevertheless, via adopting different, yet combinatory, thresholds (i)-(iii), we can draw meaningful conclusions concerning the (comparative) value of the instruments at an agent's disposal.

(i) the depth of past experience, e.g., (d11);
(ii) the minimum amount of witnesses, e.g., (d23);
(iii) the minimum rate of success, e.g., (d28).

Furthermore, we also saw that a 'best' φ-instrument Δ is not necessarily a good$_n$ φ-instrument if the success ratio expressed by threshold n is not met by Δ. Alternatively, one could consider defining 'best' in terms of ordering only those instruments that meet the threshold for qualifying as a good instrument. We leave such considerations for future work.

5.3 Defeasibility of Instrumentality Judgments

Instrumentality, as defined in this work, is a defeasible notion in three ways. First, depending on the length of the interval considered to evaluate the past, an instrument Δ may fail to qualify as an (excellent) φ-instrument once the interval is shortened.

Furthermore, it can be easily checked that the excellence of instruments may also fail to be preserved when the interval is extended.

Second, concerning the future, an instrument may fail to remain classified as an (excellent) instrument, either because an agent changes her expectations or, in the case of excellent candidate instruments, because the instrument has failed to produce the desired end in the meantime. Nevertheless, plain candidate instrumentality is preserved over time when experience is accumulated. That is, once an instrument proves to be a candidate instrument, it remains a candidate instrument.

The third and foremost defeasible aspect of our formalised instrumentality relations arises through the use of expectations. Namely, the conjectural element of instrumentality judgments imposes a defeasible characteristic on such judgments: although universal statements can be objectively true for the past, such statements remain uncertain with respect to the future. The defeasibility of expectations consists in the possibility that, although an agent α expects that the (excellent) instrument will serve its intended end once again at the moment of evaluation w, the actual successor of w is such that that instrument fails to deliver the purpose. These cases reveal a discrepancy between α's expectations and the actual future. Defeasibility with respect to the actual future is demonstrated by the fact that the formula below is not a theorem of TLAE.

$$[\Delta^{\alpha}]_{n}^{ex-instr}\varphi \rightarrow [\mathsf{A}](t(\Delta^{\alpha}) \rightarrow \varphi)$$

We close this section with a clarification concerning expectations: the conjectural element refers to the agent's expectations *at the moment of evaluation, which is what an agent expects with respect to the immediate future of that moment.* By contrast, in evaluating the past, we must ignore the agent's past expectations in selecting the agent's relevant experience. In fact, a series of unexpected events in the past may have led the agent to the conviction that a particular instrument is suitable for a specific purpose. That is, the agent may learn about instruments through unexpected events. For our formal rendering of this point, see (d16) and (d17). For instance, suppose that the moment of evaluation w is an immediate successor of a moment w': then, it might be the case that at w', the agent did not expect w to obtain. Regardless, once w obtains, the agent gains new experience concerning instrumentality judgments, as she always does whenever time passes.

6 Closing Remarks

First, we point out that we did not define von Wright's notion of *bad* instruments [38, p. 23, p. 35]. An available instrument (possibly a good or excellent instrument) is qualified as bad whenever a (side-)effect of that instrument would be undesirable. For example, whereas negotiations may be a good instrument for ending a war, using an atomic bomb may even be excellent. Still, the latter is a bad instrument because it will additionally lead to the destruction of the planet and the death of many. Here, the label 'bad' is assigned to the instrument based on its (additional) consequences (see [38] for a discussion). These may, for instance, be certain moral or social values that are violated.

Second, there is the potential to provide *deontic extensions* of the logic TLAE. For instance, deontic concepts such as 'obligation' and 'prohibition' can be incorporated in TLAE through violation constants, e.g., an action (outcome) is obligatory if and only if that action's (outcome's) complement entails a violation. This approach is known as Anderson's reduction of deontic logic [1]. Such constants denote that the agent is in a violation state. A deontic extension of the basic LAE logic from [7] was developed in [8]. In that system, various instrumentality notions concerning obligations and prohibitions were introduced. For instance, in addition to the traditional distinction between 'ought to be' and 'ought to do'—respectively, obligations about states of affairs and obligations about actions—the involvement of instrumentality statements allows for a third category called 'norms of instrumentality'. These are norms that oblige or prohibit a particular action as a means to achieving a particular end. To give an example, the norm 'It is prohibited to use nonpublic information as an instrument to acquire financial profit on the stock market' (known as the law on 'insider trading') forbids the use of such information only as a *means* to attain financial profit. The only temporal operator employed in [8] is the immediate successor modality.

Combining the above two observations, one can extend the present work to a deontic setting which allows for reasoning with prohibitions that forbid those instruments that have deontically 'bad' consequences (e.g., violations or sanctions) despite being 'good' instruments. Another direction would be to define obligatory actions in terms of those 'good' instruments (or best) that will secure an obligatory outcome. Here one can think of acquiring different notions depending on whether the agent's expectations will be involved.

Acknowledgements Kees van Berkel was supported by the projects WWTF (MA16-028) and FWF (W1255-N23). Tim S. Lyon was supported by the European Research Council (ERC) Consolidator Grant 771779 (DeciGUT). Matteo Pascucci was supported by the Štefan Schwarz Fund (2020-2023) for the project "A fine-grained analysis of Hohfeldian concepts" and by VEGA grant n. 2/0125/22 "Responsibility and modal logic". Section 2 and Section 5 are mainly due to Kees van Berkel, whereas Section 4 is mainly due to Tim S. Lyon and Matteo Pascucci. The other Sections are to be equally attributed to all the authors.

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak Republic.

References

1. Anderson, A. R., & Moore, O. K. (1957). The formal analysis of normative concepts. *American Sociological Review, 22*(1), 9–17.
2. Anscombe, G. E. M. (2000). *Intention*. Harvard University Press.
3. Åqvist, L. (2002). Old foundations for the logic of agency and action. *Studia Logica, 72*(3), 313–338.

4. Audi, R. (1989). *Practical Reasoning*. Routledge.
5. Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the Future. Agents and Choices in our Indeterminist World*. Oxford: Oxford University Press.
6. van Berkel, K. (2023). A Logical Analysis of Normative Reasoning: Agency, Action, and Argumentation, PhD dissertation, TU Wien.
7. van Berkel, K. & Pascucci, M. (2018). Notions of instrumentality in agency logic. In: Proceedings of PRIMA 2018, Springer Cham. pp. 403–419.
8. van Berkel, K., Lyon, T., & Olivieri, F. (2020). A decidable multi-agent logic for reasoning about actions, instruments, and norms. In: International Conference on Logic and Argumentation, Springer Cham. pp. 219–241.
9. Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal Logic*. Cambridge: Cambridge University Press.
10. Boudou, J., & Lorini, E. (2018). Concurrent game structures for temporal STIT Logic. *Proceedings of AAMAS, 2018*, 381–389.
11. Broersen, J. (2003). Modal Action Logics for Reasoning about Reactive Systems, PhD dissertation, Vrije Universiteit Amsterdam.
12. Broersen, J. (2011). Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic, 9*(2), 137–152.
13. Broersen, J. (2011). Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic, 40*(4), 499–530.
14. Brown, M. A. (1988). On the logic of ability. *Journal of Philosophical Logic, 17*(1), 1–26.
15. Clarke, D.S. (1987). *Practical Inferences*. Routledge Kegan & Paul.
16. Condoravdi, C., & Lauer, S. (2016). Anankastic conditionals are just conditionals. *Semantics & Pragmatics, 9*, 1–69.
17. Davidson, D. (2016). I. Agency. In R. Binkley, R. Bronaugh and A. Marras (Eds.), *Agent, Action, and Reason*, pp. 1–37. University of Toronto Press.
18. Fischer, M., & Ladner, R. (1979). Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences, 18*(2), 194–211.
19. Goldman, A. (1970). *Theory of Human Action*. Princeton: Princeton University Press.
20. Hare, R. M. (1971). *Practical Inferences*. University of California Press.
21. Herzig, A., & Lorini, E. (2010). A dynamic logic of agency I: STIT, capabilities and powers. *Journal of Logic, Language and Information, 19*(1), 89–121.
22. Horty, J., & Belnap, N. (1995). The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic, 24*(6), 583–644.
23. Hume, D. (1739). *A Treatise of Human Nature*. Oxford: Oxford University Press.
24. Lewinski, M. (2017). Practical argumentation as reasoned advocacy. *Informal Logic, 37*(2), 85–113.
25. Lorini, E., & Schwarzentruber, F. (2017). A path in the jungle of logics for multi-agent system: on the relation between general game-playing logics and seeing-to-it-that logics. *Proceedings of AAMAS, 2017*, 687–695.
26. Meyer, J. J. Ch., Broersen, J., & Herzig, A. (2015). BDI Logics. In: H. van Ditmarsch, J.Y. Halpern, W. van der Hoek, and B. Kooi (Eds.), *Handbook of Logics of Knowledge and Belief.* College Publications, pp. 453–498.
27. Rao, A. S., & Georgeff, M. P. (1995). BDI agents: from theory to practice. In: V. Lesser and L. Gasser (Eds.), *ICMAS-95, Proceedings of the first international conference of multiagent systems*. Vol. 95, pp. 312–319.
28. Raz, J. (1978). *Practical Reasoning*. Oxford University Press.
29. Sæbø, K. J. (2001). Necessary conditions in a natural language, In: C. Fery and W. Sternefeld (Eds.), *Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow*, pp. 427–449.
30. Segerberg, K. (1992). Getting started: beginnings in the logic of action. *Studia Logica, 51*(3), 347–378.
31. von Stechow, A., Krasikova, S., & Penka, D. (2006). Anankastic conditionals again. In: A Festschrift for Kjell Johan Sæbø: In Partial Fulfillment of the Requirements for the Celebration of his 50th Birthday, pp. 151–171.
32. Stoutland, F. (2010). Von Wright. In: O'Connor, T., & Sandis, C. (Eds.). *A Companion to the Philosophy of Action*, pp. 589–598
33. Walton, D. (2007). Evaluating practical reasoning. *Synthese, 157*(2), 197–240.

34. von Wright, G. H. (1957). *The Logical Problem of Induction*. New York: Barnes & Noble.
35. von Wright, G. H. (1957). Norm and Action: A Logical Enquiry. Routledge & Kegan Paul, London and Henley. Fourth impression.
36. von Wright, G. H. (1963). Practical inference. *The Philosophical Review, 72*(2), 159–179.
37. von Wright, G. H. (1968). *An Essay in Deontic Logic and the General Theory of Action*. Amsterdam: North Holland Publishing Company.
38. von Wright, G. H. (1972). *The Varieties of Goodness*. Routledge & Kegan Paul, London and Henley. Fourth impression.
39. von Wright, G. H. (1972). On so-called practical inference. *Acta Sociologica, 15*(1), 39–53.
40. Xu, M. (2010). Combinations of Stit and actions. *Journal of Logic, Language and Information, 19*, 485–503.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.