Some clarifications in logics of agency

Nicolas Troquard
Institut de Recherche en Informatique (Toulouse, France)
Laboratory of Applied Ontology (Trento, Italy)
troquard@irit.fr

Abstract. We review the logic of "seeing to it that" (STIT). We propose two new primitive operators that allow to characterize syntactically the operators Chellas', deliberative and achievement stit but also Chellas' original operator of agency $\Delta_a \varphi$. We show how it highlights their relationship and reveal differences. In particular, we remark that Chellas' stit is not an accurate simulation of Chellas' $\Delta_a \varphi$.

1 Introduction

Recently, the STIT theory has gained interest in the field of logics for computer science and artificial intelligence [Wan06, BHT06] and in ontology [TTV06, Gar06]. It is worth noting that it will be central in the introductory course "Logics of Agency and Multi-Agent systems" of ESSLLI 2007.

STIT originates from philosophy. Probably the first paper to refer to the logic of *seeing to it that* (or *theory of agents and choices in branching time*) is [BP88]. It analyzes linguistically the needs for a general theory of "an agent making a choice among alternatives that lead to an action". The thesis is that the best way to meet this goal is to augment the language with a class of sentences. The proposed class is the one of sentences of the form "Ishmael sees to it that Ishmael sails on board the Pequod" paraphrasing the sentence "Ishmael sails on board the Pequod". Thus, from any sentence describing a concrete action of an *agent a* (e.g., *sailing*) we can reformulate it into an agentive one stating that *a* sees to it that a *state of affairs* φ holds, formally: $[a\,stit\colon \varphi]$.

Formal models are provided, that constrain those of the oldest semantics for a logic of action introduced by Chellas in [Che69], such that time is linear to the past. Several agents with independent choices are also assumed. However, Belnap et al. release the assumption of discreteness.

It is important to remark that models are influenced by the observation that in a branching time framework, future-tensed statements are ambiguous to evaluate if not impossible. In general, in branching time, a moment

Proceedings of the Twelfth ESSLLI Student Session
Ville Nurmi and Dmitry Sustretov (editors)

alone does not provide enough information to determine the truth value of a sentence about the future. Prior [Pri67] and Thomason [Tho70] hence proposed to evaluate future-tensed sentences with respect to a moment *and* a particular course of time running through it. This is why states of the world in STIT models consist of 'fragmentized' moments: a moment splits up into as much indexes as there are courses of time running through it.

[BP88] is a roadmap towards a very rich theory of agency compiled in [BPX01] and [Hor01]. One of the core ideas is to capture a notion of responsibility of the agent a for the actual truth of a proposition p.

In this note, we review STIT theory (Section 2) and propose two new primitive operators that allow to characterize syntactically the operators Chellas', deliberative and achievement stit (Section 4) but also Chellas' original operator of agency $\Delta_a\varphi$. We show how it highlights their relationship and reveal differences. In particular, we remark in Section 5 that Chellas' stit is not the more accurate simulation of Chellas' original proposal $\Delta_a\varphi$ [Che69, Che92]. A brief preliminary investigation of duration of agents' activities is given in Section 6, and we conclude in Section 7.

2 The theory of agents and choices in branching time

We present here the semantics provided by Horty and Belnap [HB95].

It is embedded in the branching time framework. It is based on structures of the form $\langle W, <\rangle$, in which W is a nonempty set of moments, and $<$ is a tree-like ordering of these moments.[1] A maximal set of linearly ordered moments from W is a *history*. Thus, $w \in h$ denotes that the moment w is *on* the history h. We define $Hist$ as the set of all histories of a STIT structure. $H_w = \{h | h \in Hist, w \in h\}$ denotes the set of histories passing through w. An *index* is a pair w/h, consisting of a moment w and a history h from H_w (i.e., a history and a moment in that history). Because of branching, two different moments can lie at a same *instant*. In the following $\mathcal{A}gt$ is a nonempty set of agents and $\mathcal{A}tm$ is a set of atomic propositions. A *STIT-model* is a tuple $\mathcal{M} = \langle W, <, Choice, Instant, v\rangle$, where:

- $\langle W, <\rangle$ is a branching time structure;
- $Choice : \mathcal{A}gt \times W \to 2^{2^{Hist}}$ is a function mapping each agent and each moment w into a partition of H_w. The equivalence classes belonging to $Choice_a^w$ can be thought of as possible choices or actions available to a at w. Given a history $h \in H_w$, $Choice_a^w(h)$ represents the particular

[1] For any w_1, w_2 and w_3 in W, if $w_1 < w_3$ and $w_2 < w_3$, then either $w_1 = w_2$ or $w_1 < w_2$ or $w_2 < w_1$.

choice from $Choice_a^w$ containing h, or in other words, the particular action performed by a at the index w/h. We must have $Choice_a^w \neq \emptyset$ and $Q \neq \emptyset$ for every $Q \in Choice_a^w$;

- $Instant : W \to 2^W$: maps each moment to the set of moments lying in the same instant. It may be seen as a partition "by layers" of W into equivalence classes;
- v is valuation function $v : \mathcal{A}tm \to 2^{W \times Hist}$.

Those models are originally called $BT + I + AC$ structures, explicitly listing their main characteristics, viz. branching time, instants, agents and choices.

In STIT-models, moments may have several valuations, depending on the histories passing through them. Thus, at any specific moment, we have different valuations corresponding to the results of the different (non-deterministic) actions possibly taken at that moment.

A formula is evaluated with respect to a model and an index. Here are basic truth conditions:

$$\begin{array}{lll} \mathcal{M}, w/h \models p & \iff & w/h \in v(p), p \in \mathcal{A}tm. \\ \mathcal{M}, w/h \models \neg\varphi & \iff & \mathcal{M}, w/h \not\models \varphi \\ \mathcal{M}, w/h \models \varphi \vee \psi & \iff & \mathcal{M}, w/h \models \varphi \text{ or } \mathcal{M}, w/h \models \psi \end{array}$$

Historical necessity (or inevitability) at a moment w in a history is defined as truth in all histories passing through w:

$\mathcal{M}, w/h \models \Box\varphi \iff \mathcal{M}, w/h' \models \varphi, \forall h' \in H_w$.

There are several operators in the STIT theory. The so-called *achievement* stit was first introduced. Then Horty simplified the logic by introducing a *deliberative* one, which is deprived of the temporal aspect featured by instants, and corresponds to the previous proposition of von Kutschera [vK86, HB95]. We also present the widely used and simpler Chellas' stit:

Definition 1 (Choice equivalence). *Two moments w_1 and w_2 are $Choice_a^w$ – equivalent if (1) $instant(w_1) = instant(w_2)$ (2) w is a moment prior to both w_1 and w_2 (called* witness *moment) (3) w_1 and w_2 lie on histories belonging to the same $Choice_a^w$ partition.*

$\mathcal{M}, w/h \models [a\,astit\colon \varphi] \iff$ there is a moment $w_1 < w$ such that (for all moment w_2, $Choice_a^{w_1}$ *–equivalent* to w, $\mathcal{M}, w_2/h' \models \varphi$ for all $h' \in H_{w_2}$) and (there is some moment $w_3 \in instant(w)$ such that $w < w_3$ and $\mathcal{M}, w_3/h'' \not\models \varphi$ for some $h'' \in H_{w_3}$)

$[a\,astit\colon \varphi]$ means that agent a has ensured that φ holds now by making a choice previously, and if he had made a different choice, φ could have been false at the present instant.

$$\begin{array}{lcl} \mathcal{M}, w/h \models [a\, dstit\colon \varphi] & \iff & \forall h' \in Choice_a^w(h), \mathcal{M}, w/h' \models \varphi \text{ and} \\ & & \exists h'' \in H_w, \mathcal{M}, w/h'' \not\models \varphi \\ \mathcal{M}, w/h \models [a\, cstit\colon \varphi] & \iff & \forall h' \in Choice_a^w(h), \mathcal{M}, w/h' \models \varphi \end{array}$$

Intuitively $[a\, cstit : \varphi]$ means that agent a's current choice ensures φ, whatever the other agents do. $[a\, dstit : \varphi]$ adds the fact that φ was not settled, so, in a sense, that agent a is responsible for φ. Truth conditions of those operators do not depend on instants. They can be evaluated in simpler models called $BT + AC$ structures.

3 A discrete time framework

What can be now of interest, is to understand the underlying link between the three main versions of STIT operator, viz. Chellas' stit, deliberative stit and achievement stit. The deliberative stit can be defined from Chellas' plus historical necessity since the following holds:

$$[a\, dstit\colon \varphi] \leftrightarrow [a\, cstit\colon \varphi] \wedge \neg\Box\varphi$$

The other way round, we have $[a\, cstit : \varphi] \leftrightarrow [a\, dstit : \varphi] \vee \Box\varphi$. The link between deliberative and Chellas' stit is then quite obvious. However, a formal link of the achievement stit with them is more involved. We nevertheless claim that, because of the complex semantics of $[\_ astit : \_]$, such a relationship can provide a neat picture of the fundamental aspects of the theory of choice in time. And in order to stick to the *fundamental* aspects, let us first simplify the framework by some usual assumptions. They at least are usual in a discipline like computer science, and have the merit to rule out some features that were enabled just for a matter of generality, and thus unfortunately hid some other essential features. Belnap and colleagues refrained from taking position on the nature of time.

> "[...] no assumption whatsoever is made about the order type that all histories share with each other and with *Instant*. For this reason the present theory of agency is immediately applicable regardless of whether we picture succession as discrete, dense, continuous, well-ordered, some mixture of these, or whatever; and regardless of whether histories are finite or infinite in one direction or the other." ([BPX01, p. 196].)

We thus consider the assumption of time isomorphic to the set of natural numbers interesting to study. We would like to investigate how such a simplification can strengthen our understanding of logics of agency. We explicit discreteness as follows:

Definition 2. *The total function instantof* $: W \to \mathbb{N}$ *associates an instant to each moment. The function atinstant* $: \mathbb{N} \to 2^W$ *associates each instant to the set of moments lying in.*

Hypothesis 1. *Histories are regular: (1)* $\forall h, h' \in Hist,\ \forall w \in h,\ \exists w' \in h',\ s.t.\ instantof(w) = instantof(w')$ *(2) for some* $h \in Hist$ *and* $w \in h$*, if* $instanof(w) = i$ *then* $\forall j < i,\ \exists w' \in h\ s.t.\ instantof(w') = j$.

Moreover, we assume the existence of a root:

Hypothesis 2. *There is a moment* w *such that there is no* w' *such that* $w' < w$.

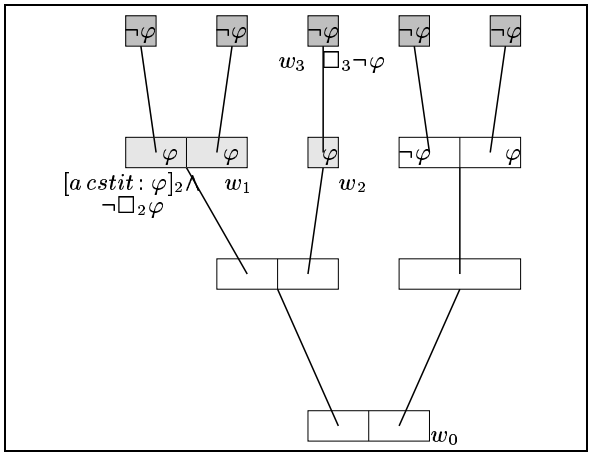

Figure 1.1: (Time goes upward.) At w_0, a can make the choice that φ is true in two steps, even though it is not settled it will be true at that instant. At w_1 (or w_2) it will be the case $[a\,astit: \varphi]$. Indeed, for some $h \in w_1$, $w_1/h \models [a\,cstit: \varphi]_2$ (φ is true at every index of w_1 and w_2) and $w_1/h \models \neg\Box_2\varphi$. At w_0 it is however already settled that in three steps, φ will be false: for some $h' \in w_3$, $w_3/h' \models \Box_3\neg\varphi$. ($\varphi$ is true at every (upper) dark grey moment.)

4 $\mathbb{N}$STIT

In order not to get confused let us call $\mathbb{N}$STIT the logic interpreted by $BT + I + AC$ structures constrained by the hypothesis previously presented, and syntactically extending the STIT theory presented in Section 2 (with a language containing operators Chellas', deliberative and achievement stit) with the two following collections of operators indexed by a natural number k:

- $\mathcal{M}, w/h \models \Box_k\varphi \iff \exists w_0 \leq w,\ instantof(w_0) = instantof(w) - k,\ \forall w' \in Instant(w) \cap (\bigcup_{h' \in H_{w_0}} h'),\ \forall h' \in w',\ \mathcal{M}, w'/h' \models \varphi$
  It reads that "k instants ago, it was settled that φ would be true now".

- $\mathcal{M}, w/h \models [a\, cstit\colon \varphi]_k \iff \exists w_0 \leq w,\ instantof(w_0) = instantof(w) - k,\ \forall w'\ Choice_a^{w_0} - equivalent$ of $w,\ \forall h' \in w',\ \mathcal{M}, w'/h' \models \varphi$
  It reads that "k instants ago, agent a ensured that φ would be true now".

Analogously to the achievement stit, we call w_0 in the previous truth conditions the *witness moment* of $[a\, cstit\colon \varphi]_k$ or $\Box_k\varphi$.

We offer to $\mathbb{N}$STIT a mechanism close to what exists in Hybrid Logic [BdRV01, Chap. 7]. We assume the existence of a set $\{0, 1, \ldots\}$ of specific atomic formulae that we could call *nominals*. We thus constrain the models such that $\mathcal{M}, w/h \models i$ iff $instant(w) = i$. Our account is nevertheless different from Hybrid Logic since genuine nominals should be true at exactly one moment/history pair. (See for example [BG01] for a concrete account of hybrid temporal logic.)

Now, let us exhibit some interesting validities, candidates to the status of axioms for future developments.

(NP) $\quad 0 \to \neg\Box_1\top$

(P) $\quad 0 \vee \Box_1\top$

(Mon) $\quad \Box_{k+1}\top \to \Box_k\top$

(Sett-1) $\quad \Box_k k \to \Box_k\top$

(Sett-2) $\quad k \leftrightarrow \Box_k k$

(NP) captures that there is no past beyond the instant 0. (P) on the contrary means that whenever we do not stand at instant 0 we can 'step back' in the temporal structure. (Mon) means says that when we can look back at $k+1$ steps, we can look back at k steps as well. (Sett-1) says that k times ago, it was settled that we would be standing at instant k only if we can look back at k steps. (Sett-2) means that we are standing at instant k iff it was already settled k steps ago that we would stand at instant k now.

We are now ready to see how the operators of the STIT language relate to our new primitives.

Proposition 1. *The four following formulae are valid:*

- $\Box\varphi \leftrightarrow \Box_0\varphi$
- $[a\, cstit\colon \varphi] \leftrightarrow [a\, cstit\colon \varphi]_0$
- $[a\, dstit\colon \varphi] \leftrightarrow [a\, cstit\colon \varphi]_0 \wedge \neg\Box_0\varphi$
- $i \to ([a\, astit\colon \varphi] \leftrightarrow \bigvee_{k=1}^{i}([a\, cstit\colon \varphi]_k \wedge \neg\Box_k\varphi))$

From the last item, we can have a local definition of achievement stit for every instant. It is indeed similar to the definition of tense operator 'until' and 'since'. (See [BG01, Sect. 4.1].) Historical necessity, Chellas' stit and deliberative stit on the other hand, can be completely defined from our new primitives.

Instances of the new primitive operator of agency are intrinsically related and obey the following property:

Proposition 2. $[a\,cstit\colon \varphi]_{k_1} \to [a\,cstit\colon \varphi]_{k_2}$, *for every* $k_2 < k_1$.

5 Comments on Chellas' $\Delta_a\varphi$

In [Che92], Brian Chellas turns back to his operator of agency introduced in [Che69]. As in theories of agents and choices in branching time, truth values of the language are in terms of times (alias instants), histories and agents, plus *certain* relations. Here, we quickly show how we can define $\Delta_a\varphi$ fairly in $\mathbb{N}\mathsf{STIT}$, and also suggest that Chellas' stit operator does not match perfectly.

5.1 Semantics of time and actional alternatives

The set of times is taken to be the set of integers. We write $t < t'$ to state that t is earlier than t' and $t \leq t'$ to state it is not later. Histories are functions from times to states of affairs (alias moments), and $h(t)$ represents the state of affairs in history h at time t. Two time-indexed relations between histories are then defined. $h =_t h'$ means that histories h and h' have the same past at time t; $h \equiv_t h'$ means that they share the same past *and* the same present. Formally,

- $h =_t h'$ iff $h(t') = h'(t')$ at every time $t' < t$
- $h \equiv_t h'$ iff $h(t') = h'(t')$ at every time $t' \leq t$

Given a state of affairs h_t, Chellas uses the term *future cone* for the set of histories emanating from h_t. Two histories are in the future cone of $h(t)$ if $h \equiv_t h'$.

Instigative alternatives. The relation $R^a_t(h, h')$ is used to mean that h' is an *instigative alternative* of h for agent a at time t. The relation is reflexive and $R^a_t(h, h')$ only if $h =_t h'$. Instigative alternatives capture the idea of histories "under the control" or "responsive to the action" of a at t.

Truth conditions of the operator of agency is given by:

$$(h, t) \models \Delta_a\varphi \iff (h', t) \models \varphi,\ \forall h' \text{ s.t } R^a_t(h, h')$$

5.2 Chellas' stit is not $\Delta_a\varphi$

In addition to our short overview, it is interesting and helpful to consider Krister Segerberg's interpretation of the operator in [Seg92]. Segerberg calls $R_t^a(h, h')$ the cone of 'actional alternatives' and observes that in the truth value of $\Delta_a\varphi$, "the cone Chellas wishes to consider has its apex at the immediately preceding time". This is indeed a consequence of the constraint that two histories h and h' are instigative alternatives only if $h =_t h'$.

Finally, we can define more appropriately the operator in $\mathbb{N}\mathsf{STIT}$ as follows:

$$\Delta_a\varphi \triangleq [a\,cstit\colon \varphi]_1$$

It thus clearly differs from $[a\,cstit\colon \varphi]$ which we remind is logically equivalent in $\mathbb{N}\mathsf{STIT}$ to $[a\,cstit\colon \varphi]_0$. There is a temporal switch between them. One must be aware of a possible misconception of the Chellas' stit, since it does not reflect Chellas' original operator. If Chellas had in mind something similar to Chellas' stit when he made up his $\Delta_a\varphi$ operator, he would have constrained the instigative alternatives (or actional alternatives) such that $R_t^a(h, h')$ only if $h \equiv_t h'$.

Still, it does not mean that $[a\,cstit\colon \varphi]_1$ *is* $\Delta_a\varphi$ without nuance. Our definition also suffers the fact that Chellas did not impose a "future branching only" [Che92, p. 489] nature of time and the independence of agents, while we inherit them from $BT + AC$ structures.

6 Duration of an activity

Now, those operators indexed with a natural number k may seem odd. But this is not odder than an iterated operator 'next' permitting to jump from instant to instant along a history. This is actually interesting to see what is going on if we allow such an operator:

$$\mathcal{M}, w/h \models \mathbf{X}\varphi \iff \exists w\ w < w',\ \nexists w''\ w < w'' < w',\ s.t.\ \mathcal{M}, w'/h \models \varphi$$

In order to highlight how our primitive operators behave over time, it is easy to prove that $[a\,cstit\colon \mathbf{X}^k\varphi] \leftrightarrow \mathbf{X}^k[a\,cstit\colon \varphi]_k$, and $\Box\mathbf{X}^k\varphi \leftrightarrow \mathbf{X}^k\Box_k\varphi$.

Let us designate a *chain* as being a set of linearly ordered moments. "In branching time, chains represent certain complex concrete events" [BPX01, p. 181].

While in the original STIT theory the $[\_\,cstit\colon \_]$ permits to express that an agent selects some set of histories (*unbounded* sets of ordered moments), underlying events are loosely characterized: they correspond to every chain we can construct on those histories. With $[\_\,cstit\colon \_]_k$ we clearly identify the set of events the agent has brought about: events composed of moments

between the moment of choice w and moments that are on the selected histories not farer than k instants after w. We see that as a strength of the language.

An example. To give some intuition of possible applications of $[a\,cstit:\varphi]_k$ consider the following example. In an institutional context, it can be useful to reason about the length of an activity. For instance, given an operator for obligation $\bigcirc$, we could have a formula like

$$phd(Mary) \rightarrow \bigcirc[Mary\,cstit\colon Mary\_has\_written\_her\_thesis]_{24}$$

in the domain description, to state that a student can obtain a PhD only if he or she has achieved the writing of the thesis and has spent *at least* 24 months working on it.[2] From Proposition 2, it indeed captures the notion of minimum. In such a modeling, it is like Mary chose at least 24 'clock ticks' ago (that happen here to correspond to months) to write a thesis and it happens to have succeeded *now*.

7 Concluding remarks

The contribution here is humble: make clearer the link between logical operators by adding what can be seen conceptually harmless constraints in a discipline like computer science. First, it clearly highlights that the deliberative stit is a *local* achievement stit, or an achievement stit having the current moment as a witness. Second, it permits us to provide a more appropriate simulation of Chellas' original operator of agency which was simply impossible without assuming discreteness.

Acknowledgment

I am debtful to Laure Vieu for her crucial observations on $\mathbb{N}$STIT and to Jan Broersen for pointing to me a problem in a preliminary version. I would also like to thank anonymous reviewers of this ESSLLI Student Session who did particularly relevant remarks regarding this paper.

Bibliography

[BdRV01] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic.* Cambridge University Press, 2001.

[BG01] Patrick Blackburn and Valentin Goranko. Hybrid Ockhamist Temporal Logic. In Bettini, C. and Montanari, A, editor, *Proceedings of the 8th Int.*

[2] In France a minimum of 2 years is imposed.

Symp. on Temporal Representation and Reasoning (TIME-01), pages 183–188. IEEE Computer Society Press, 2001.

[BHT06] Jan Broersen, Andreas Herzig, and Nicolas Troquard. Embedding Alternating-time Temporal Logic in strategic STIT logic of agency. *Journal of Logic and Computation*, 16(5):559–578, 2006.

[BP88] Nuel Belnap and Michael Perloff. Seeing to it that: a canonical form for agentives. *Theoria*, 54:175–199, 1988.

[BPX01] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world.* Oxford, 2001.

[Che69] Brian Chellas. *The Logical Form of Imperatives.* PhD thesis, Philosophy Department, Stanford University, 1969.

[Che92] Brian F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3/4):485–518, 1992.

[Gar06] Pawel Garbacz. The Instrumental Stit A Study of Action and Instrument. In Brandon Bennett and Christiane Felbaum, editors, *International Conference on Formal Ontology in Information Systems, Baltimore, Maryland, USA*, pages 167–178. IOS Press, 2006.

[HB95] John F. Horty and Nuel D. Belnap, Jr. The deliberative STIT: A study of action, omission, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

[Hor01] John F. Horty. *Agency and Deontic Logic.* Oxford University Press, Oxford, 2001.

[Pri67] A.N. Prior. *Past, Present, and Future.* Clarendon Press, 1967.

[Seg92] Krister Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51(3/4):347–378, 1992.

[Tho70] Richmond Thomason. Indeterminist time and truth-value gaps. *Theoria*, 36:264–81, 1970.

[TTV06] Nicolas Troquard, Robert Trypuz, and Laure Vieu. Towards an ontology of agency and action : From STIT to OntoSTIT+. In Brandon Bennett and Christiane Felbaum, editors, *International Conference on Formal Ontology in Information Systems, Baltimore, Maryland, USA*, pages 179–190. IOS Press, 2006.

[vK86] Franz von Kutschera. Bewirken. *Erkenntnis*, 24(3):253–281, 1986.

[Wan06] Heinrich Wansing. Tableaux for multi-agent deliberative-stit logic. In Guido Governatori, Ian Hodkinson, and Yde Venema, editors, *Advances in Modal Logic, Volume 6*, pages 503–520. King's College Publications, 2006.