A STIT-Extension of ATL

Jan Broersen[1], Andreas Herzig[2], and Nicolas Troquard[2]

[1] Department of Information and Computing Sciences, Utrecht University
[2] Institut de Recherche en Informatique de Toulouse

Abstract. A problem in many formalisms for reasoning about multi-agent systems, like ATL or PDL, is the inability to express that a certain complex action (as in PDL), choice or strategy (as in ATL) is performed by an agent. However, in so called STIT-logics, this is exactly the main operator: seeing to it that a certain condition is achieved. Here we present an extension of ATL, introducing ideas from STIT-theory, that can express that a group of agents A perform a certain strategy. As a demonstration of the applicability of the formalism, we show how it sheds new light on the problem of modelling 'uniform strategies' in epistemic versions of ATL.

1 Introduction

The present paper introduces a so called 'strategic STIT-operator' in the framework of ATL [1, 2]. For those unfamiliar with the STIT-framework: the characters 'STIT' are an acronym for 'seeing to it that'. STIT logics [3, 4, 5] originate in philosophy, and can be described as endogenous logics of agency, that is, logics of agentive action where actions are not made explicit in the object language. To be more precise, expressions $[A \mathit{stit} : \varphi]$ of STIT-logic stand for 'agents A see to it that φ', where φ is a (possibly) temporal formula. The main virtue of STIT logics is that, unlike most (if not all) other logical formalisms, they can express that a choice or action is actually performed / taken / executed by an agent. The aim of the present paper is thus to add this type of expressivity to the ATL-framework. But not only do we want to add the standard STIT expressivity, we intend to define a *strategic* version of STIT as an addition to ATL. This enables us to express what it means that a group of agents performs / takes / executes a certain *strategy*. ATL itself can only talk about the *existence* or 'availability' of certain strategies, not that they are actually being performed.

We consider the definition of a semantics for a strategic version of STIT within the ATL-framework as the main contribution of this paper. Indeed, within the community working on the STIT framework of Belnap [3, 4, 5] and Horty [6], it is perceived as an open problem how to define a suitable notion of *strategic* STIT. As a corollary the semantics shows how we can make the implicit quantifications in the semantics of the ATL operators explicit in the object language: the two central ATL operators will each be decomposed into a strategy quantifier and a strategic STIT operator. To demonstrate the applicability of the formalism, in section 4 we will discuss an extension with epistemic notions, and discuss the problem of 'uniform strategies'. This has also been the subject of [7], but section 4 adds in some new insights. Also the present paper differs

M. Fisher et al. (Eds.): JELIA 2006, LNAI 4160, pp. 69–81, 2006.

from [7] in that we introduce epistemic notions in a STIT-extension of the ATL framework, whereas [7] introduces epistemic notions in the STIT framework. Furthermore, here we deal with general strategies, where [7] only deals with one-step strategies.

Central to our approach will be to evaluate ATL-STIT formulas with respect to strategy / state pairs. Tinkering with the units of evaluation has been suggested before in the literature on ATL and STIT. Horty [6] indeed already suggests it to define a notion of strategic STIT. Although Horty suggests two possible approaches, he circumvents the problem of actually giving definitions for the strategic STIT by syntactically disallowing this operator to occur without an attached operator quantifying over histories. Müller [8] suggests evaluation with respect to strategies to deal with the notion of continuous action within the STIT framework, and Jamroga and gnotes [9] suggest to evaluate with respect to sets of worlds to solve the problem of uniform strategies in epistemic ATL. We will discuss these related approaches in more detail in section 5.

In earlier work [10] we investigated the similarities between the ATL and STIT frameworks. The present paper is a demonstration of our opinion that there can be a fruitful exchange of techniques and ideas between both frameworks. The idea for investigating strategic versions of STIT operators originates from Belnap (Horty [6] mentions an unpublished manuscript) and Horty. Here we show how we can successfully define this concept in the ATL setting. An ensuing next step would then be to transfer these ideas back to the STIT framework.

2 The Meaning of 'Agents a Performing a Strategy'

First we need to explain that we think that 'strategy' seems not the best term for the moment-to-action-functions defined in this paper. We feel it would be more in line with established general AI terminology to call them 'tactics' or 'conditional plans'. Strategies are usually associated with choices for more abstract (sub-)goals, while tactics are indeed more concrete (conditional) plans for reaching these goals. Yet, to adhere to established terminology in both STIT theory and ATL, we will also refer to the conditional plans as 'strategies'.

An important conceptual first question is then what it exactly means to say that 'a group is performing a strategy'. Is whether or not 'a group is performing a strategy' actually a sensible concept amenable to logical truth? For instance, in what sense can it be true that 'agent j, who is still at home, presently performs the strategy of going to the railway station'? A strong intuition is that performing an action / choice is a local matter concerning the present. The problem then seems to be that at any future point j may reconsider his strategy. Half way to the railway station he may decide to go to the cafe and have a beer instead. So how could we ever say that an agent is performing a certain strategy presently if at any future point he may decide to deviate from it? Is it not that all we can say is that an agent is *committed* to a certain strategy, thereby leaving room for the possibility that an agent reconsiders his strategy?

Our answer is that the notion of commitment to a strategy actually presupposes a notion of performing a strategy. How can we say that an agent is committed to going to the railway station (which, one way or the other, expresses a certain preference for some strategies over others) if we cannot say what it means for the agent to actually

perform going to the railway station? The same holds for strategic contents of epistemic notions. For instance, if we want to say that we belief that agent A performs a certain strategy, than first we have to know what it means that A performs a strategy. So, if we do not accept 'agent A is performing a strategy for φ' as a meaningful proposition, we cannot accept 'agent A is *committed* to performing a strategy for φ' and 'agent B *beliefs* that agent A is performing a strategy for φ' as meaningful propositions either. The conclusion then is that although it is maybe strange to think about the truth of propositions talking about performance of strategies as such, it is not at all strange to reason with these propositions within the scope of motivational and epistemic modalities. Human agents do this all the time. Presently we are *committed* to performing the strategy to finish writing this paper (in time), which presupposes that we know what it means to actually perform this strategy. Also, we *believe* that president Bush is performing a strategy of world destruction, which presupposes that it is clear what it means to be performing such a strategy. So the notion of performing a strategy is not inherently problematic. We reason with the notion all the time, and the present proposal defines a semantics for it.

3 ATL-STIT

We present a STIT extension of ATL ([1, 2]) using a non-standard, but concise and intuitive syntax and semantics.

3.1 Core Syntax, Abbreviations and Intended Meanings

Definition 1. *Well-formed formulas of the temporal language $\mathcal{L}_{ATL\text{-}STIT}$ are defined by:*

$$\varphi, \psi, \ldots := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Diamond_A \varphi \mid \Box_A \varphi \mid [A]\eta \mid \langle A \rangle \eta$$
$$\eta, \theta, \ldots := \phi U^{ee} \psi$$

where $\varphi, \psi, \ldots$ represent arbitrary well-formed formulas, $\eta, \theta, \ldots$ represent temporal path formulas, the p are elements from an infinite set of propositional symbols $\mathcal{P}$, and A is a subset of a finite set of agent names E (we define $\overline{A} \equiv_{def} E \setminus A$). We use the superscript 'ee' for the until operator to denote that this is the version of 'the until' where φ is not required to hold for the present, nor for the point where ψ, i.e., the present and the point where ϕ are *both* excluded. The operators $\Box_A \varphi$ and $\Diamond_A \varphi$ are universal and existential quantifiers over strategies, respectively. The STIT operators $[A]\eta$ and $\langle A \rangle \eta$ are read as 'agents A strategically see to it that η' and 'agents A strategically allow the possibility for η', respectively. The combined operator $\Diamond_A [A]\eta$ is read as 'Agents A have a strategy that ensures η' (this is the 'classical' ATL operator, usually written as $\langle\langle A \rangle\rangle \eta$), and the dual $\Box_A \langle A \rangle \eta$ is read as 'A have no strategy to avoid that possibly η'. A more precise explanation of the intended semantics is as follows:

$\Diamond_A \varphi$: there is a strategy (for the set of agents A, from the current state) such that φ
$\Box_A \varphi$: for all strategies (of the set of agents A, from the current state) φ

The intended interpretations for the new *strategic STIT operators* are:

$[A](\varphi U^{ee}\psi)$: agents A perform a strategy that, whatever strategy is taken by agents$\overline{A}$, ensures that eventually, at some point m, the condition ψ will hold, while φ holds from the next moment until the moment before m

$\langle A\rangle(\varphi U^{ee}\psi)$: Agents A perform a strategy giving agents $\overline{A}$ the possibility to perform a strategy such that eventually, at some point m, the condition ψ will hold, while φ holds from the next moment until the moment before m

We use standard propositional abbreviations, and also define the following operators as abbreviations.

Definition 2.

$[A]X\varphi \equiv_{def} [A](\bot U^{ee}\varphi)$ $\quad$ $\langle A\rangle X\varphi \equiv_{def} \langle A\rangle(\bot U^{ee}\varphi)$

$[A]F\varphi \equiv_{def} [A](\top U^{ee}\varphi)$ $\quad$ $\langle A\rangle F\varphi \equiv_{def} \langle A\rangle(\top U^{ee}\varphi)$

$[A]G\varphi \equiv_{def} \neg\langle A\rangle F\neg\varphi$ $\quad$ $\langle A\rangle G\varphi \equiv_{def} \neg[A]F\neg\varphi$

$[A](\varphi U^{e}\psi) \equiv_{def} [A](\varphi U^{ee}(\varphi \wedge \psi))$ $\quad$ $\langle A\rangle(\varphi U^{e}\psi) \equiv_{def} \langle A\rangle(\varphi U^{ee}(\varphi \wedge \psi))$

$[A](\varphi U^{e}_{w}\psi) \equiv_{def} \neg\langle A\rangle(\neg\psi U^{e}\neg\varphi)$ $\quad$ $\langle A\rangle(\varphi U^{e}_{w}\psi) \equiv_{def} \neg[A](\neg\psi U^{e}\neg\varphi)$

The informal meanings of the formulas are as follows (the informal meanings in combination with the $\langle A\rangle$ operator follow trivially):

$[A]X\varphi$: agents A strategically ensure that at any next moment φ will hold

$[A]F\varphi$: agents A strategically ensure that eventually φ will hold

$[A]G\varphi$: agents A strategically ensure that φ holds henceforth

$[A](\varphi U^{e}\psi)$: agents A strategically ensure that, eventually, at some point the condition ψ will hold, while φ holds from the next moment until then

$[A](\varphi U^{e}_{w}\psi)$: agents A strategically ensure that, if eventually ψ will hold, then φ holds from the next moment until then, or forever otherwise

Note that all STIT formulas refer strictly to the future. Also, for instance, a formula like $[A]G\varphi$ saying that φ holds henceforth, does not imply that φ holds now.

Alternatively, we could have taken $[A]\varphi U^{e}\psi$ and $[A]G\varphi$ as the basic operators of our langauge, which would enable us to define $\langle A\rangle\varphi U^{e}\psi$ in terms of them. A similar choice appears for the definition of related logics like ATL and CTL. However, we prefer the symmetry of the present setup, and we think the semantics of the new weak STIT operator $\langle A\rangle\varphi U^{ee}\psi$ deserves a definition in terms of truth conditions.

3.2 Model Theoretic Semantics

We use alternating transition systems (ATSs) for the semantics. Goranko and Jamroga [11] argue that to define the semantics of ATL, multi-player game models (MGMs) provide more intuitive semantic representations in many examples. However, ATSs are closer to the models used for STIT logics. And actually we do not fully agree that ATSs are better than MGMs as semantic structures for ATL. We believe it is better not

to decorate semantic structures with superfluous information. For instance, in MGMs the actions have explicit names. However ATL is an endogenous temporal formalism where the strategies (which can be seen as conditional plans) are not explicit in the object language. So, ATL is not, so to say, 'aware' of the actions names. We will come back to this point in section 4.2.

The assumption behind ATSs is that agents have choices, such that the non-determinism of each choice is *only* due to the choices other agents have at the same moment. Thus, the simultaneous choice of al agents together, always brings the system to a unique follow-up state. In other words, if an agent would know what the choices of other agents would be, given his own choice, he would know exactly in which state he arrives.

Definition 3. *An ATS* $\mathcal{M} = (S, C, \pi)$*, consists of a non-empty set* S *of states, a total function* $C : E \times S \mapsto 2^{2^S}$ *yielding for each agent and each state a set of choices (informally: 'actions') under the condition that the intersection of each combination of choices for separate agents gives a unique next system state (i.e., for each* s*, the function* $RX(s) = \{\bigcap_{a \in E} Ch_a \mid Ch_a \in C(a, s)\}$ *yields a non-empty set of singleton sets representing the possible follow-up states of* s*), and, finally, an interpretation function* π *for propositional atoms.*

Note that from the condition on the function C it follows that the choices for each individual agent at a certain moment in time are a partitioning of the set of all choices possible for the total system of agents, as embodied by the relation $\mathcal{R}^{sys} = \{(s, s') \mid s \in S \text{ and } \{s'\} \in RX(s)\}$. And, also note that this latter condition does not entail the former. That is, there can be partitions of the choices for the total system that do not correspond to the choices of some agent in the system. Now we are ready to define strategies relative to ATSs.

Definition 4. *Given an ATS, a strategy* α_a *for an agent* a*, is a function* $\alpha_a : S \mapsto 2^S$ *with* $\forall s \in S : \alpha_a(s) \in C(a, s)$*, assigning choices of the agent* a *to states of the ATS.*

In semantics for ATL, strategies are often defined as mappings $\alpha_a : S^+ \mapsto 2^S$, from finite *sequences* of states to choices in the final state of a sequence. However, to interpret ATL, this is not necessary, because ATL is not expressive enough to recognize by which sequence of previous states a certain state is reached (but ATL* is). More in particular, without affecting truth of any ATL formula, we can always transform an ATS into one where $\mathcal{R}^{sys}$ is tree-like. On tree structures it is clear right away that a mapping from states to choices in that state suffices, since any state can only be reached by the actions leading to it. We come back to this point in section 4.

Definition 5. *Strategy functions* α_a *for individual agents* a *are straightforwardly combined to system strategy functions* $\alpha_E : S \times E \mapsto 2^S$ *for the full set of agents* E*. Then* $\alpha_E(s, a)$ *yields the choice of agent* a *in state* s *determined by the system strategy* α_E*. However, central to our semantics will be* partial *strategy functions* $\alpha_A : S \times E \mapsto 2^S$*, where* $A \subseteq E$*. These functions are partial in the sense that no choices are defined for the agents* $\overline{A}$*. For* $B \subseteq A$ *we use the notation* $\alpha_A{\restriction_B}$ *to denote the partial strategy function that is the restriction of the partial strategy function* α_A *to the domain of agents* B *(note*

that $\alpha_A \restriction_A = \alpha_A$*). Furthermore, for* $A \cap B = \emptyset$*, we use* $\alpha_A | \beta_B$ *to denote the minimal joined partial strategy function build from* α_A *and* β_B *such that* $(\alpha_A | \beta_B) \restriction_A = \alpha_A$ *and* $(\alpha_A | \beta_B) \restriction_B = \beta_B$*.*

As said, if in a given state all agents in the system have fixed their choice, a unique next state is determined by the intersection of all choices. Analogously, if all agents in the system have fixed a strategy, from any given point, a unique infinite path into the future is determined by the intersection of all choices in the strategies. We use this in the next definition.

Definition 6. *Given a system strategy* α_E*, we define the follow up function* $F_{\alpha_E} : S \mapsto S$ *as the intersection of all choices for individual agents, that is,* $F_{\alpha_E}(s) = \bigcap_{a \in E} \alpha_E(s, a)$*. Then, by* $(F_{\alpha_E})^n(s)$ *we denote the unique state that results from state* s *by taking* n *steps of the system strategy* α_E

Now we are ready to define the formal semantics of the language $\mathcal{L}_{\text{ATL-STIT}}$. The essential new aspect of this semantics is that it evaluates formulas with respect to strategy / state pairs. For a given fixed ATS, the set of all possible strategies for any group of agents A is well defined. So technically there is no problem with evaluation against strategy / state pairs. The pairs of an ATS form a two-dimensional modal structure, with group strategies and (impersonal) moments constituting the two 'axis'. As is customary for multi-dimensional possible world structures, we have modal operators interpreted on individual dimensions only: the strategy quantification operators $\Diamond_A \varphi$ and $\Box_A \varphi$ are interpreted on the dimension of strategies, relative to a *fixed* moment, and the temporal STIT operators $[A]\phi U^{ee} \psi$ and $\langle A \rangle \phi U^{ee} \psi$ are interpreted on the moments, relative to a *fixed* strategy.

But then the question remains: why should we *want* to evaluate against strategy / state pairs? It is clear that we want to give semantics to the strategic STIT operators. Truth of such operators cannot be determined with respect to states or moments alone, since in general, at the same moment, agents have a choice between several strategies. If we really want to give meaning to an operator that enables us to express that it is *true* that an agent, or group of agents performs a strategy, we have to take the possible strategies as units of evaluation. Then, with group strategies as abstract possible worlds, through evaluation in such worlds we can determine whether or not it is true that a group of agents strategically see to something.

Definition 7. *Validity* $\mathcal{M}, \alpha_A, s \models \varphi$*, of an ATL-STIT-formula* φ *in a strategy / state pair* (α_A, s) *of an ATS* $\mathcal{M} = (S, C, \pi)$ *is defined as:*

$$
\begin{array}{ll}
\mathcal{M}, \alpha_A, s \models p & \Leftrightarrow s \in \pi(p) \\
\mathcal{M}, \alpha_A, s \models \neg\varphi & \Leftrightarrow \textit{not } \mathcal{M}, \alpha_A, s \models \varphi \\
\mathcal{M}, \alpha_A, s \models \varphi \wedge \psi & \Leftrightarrow \mathcal{M}, \alpha_A, s \models \varphi \textit{ and } \mathcal{M}, \alpha_A, s \models \psi \\
\mathcal{M}, \alpha_A, s \models \Diamond_B \varphi & \Leftrightarrow \exists \beta_B \textit{ such that } \mathcal{M}, \beta_B, s \models \varphi \\
\mathcal{M}, \alpha_A, s \models \Box_B \varphi & \Leftrightarrow \forall \beta_B \textit{ it holds that } \mathcal{M}, \beta_B, s \models \varphi \\
\mathcal{M}, \alpha_A, s \models [B]\phi U^{ee} \psi & \Leftrightarrow \forall \beta_{\overline{A \cap B}} \textit{ it holds that } \exists n > 0 \textit{ such that} \\
 & \quad (1)\ \mathcal{M}, \alpha_A, (F_{\alpha_E})^n(s) \models \psi \textit{ and} \\
 & \quad (2)\ \forall i \textit{ with } 0 < i < n \textit{ we have } \mathcal{M}, \alpha_A, (F_{\alpha_E})^i(s) \models \varphi \\
 & \quad \textit{where } \alpha_E \textit{ is defined as: } \alpha_E = \alpha_A \restriction_{A \cap B} | \beta_{\overline{A \cap B}}
\end{array}
$$

$\mathcal{M}, \alpha_A, s \models \langle B\rangle \phi U^{ee} \psi \Leftrightarrow \exists \beta_{\overline{A\cap B}}$ *and* $\exists n > 0$ *such that*
(1) $\mathcal{M}, \alpha_A, (F_{\alpha_E})^n(s) \models \psi$ *and*
(2) $\forall i$ *with* $0 < i < n$ *we have* $\mathcal{M}, \alpha_A, (F_{\alpha_E})^i(s) \models \varphi$
where α_E *is defined as:* $\alpha_E = \alpha_A \restriction_{A\cap B} | \beta_{\overline{A\cap B}}$

Validity on an ATS $\mathcal{M}$ is defined as validity in all strategy / state pairs of the ATS. If φ is valid on an ATS $\mathcal{M}$, we say that $\mathcal{M}$ is a model for φ. General validity of a formula φ is defined as validity on all possible ATSs. The logic ATL-STIT is the subset of all general validities of $\mathcal{L}_{ATL\text{-}STIT}$ over the class of ATSs.

Note that due to the constraints on ATSs, if an atomic proposition is evaluated true on a strategy / state pair, all strategy / state pairs with the same state, will also have to evaluate to true, because for atomic propositions assignment of truth values is independent of the strategy. In Horty and Belnap's STIT formalisms atomic propositions can have different valuations at the same moment, depending on what history they are. In our setting, only formulas referring strictly to the future can evaluate to different values for the same moment, depending on the strategy with respect to which they are evaluated. We might say that in Horty's formalisms choices may affect the present, while our choices may only affect the strict future (both frameworks assume it makes no sense to account for choices affecting the past).

The most important aspect of the above definition is the truth condition for the STIT operators. Note that we evaluate the STIT operator $[B]\eta$ for a group of agents B with respect to a strategy for another group A. The truth condition expresses exactly in what sense the group B may see to it that η in a strategy of group A, namely, exactly if η is guaranteed by the agents in the intersection of both groups. This exploits the intuition that if a subgroup of agents sees to it that η, all supergroups also see to it that η. Now we show that ATL is a fragment of the logic ATL-STIT.

Theorem 1. *The logic ATL is the fragment of the logic ATL-STIT determined by the definitions* $\langle\langle A\rangle\rangle\eta \equiv_{def} \Diamond_A[A]\eta$ *and* $[[A]]\eta \equiv_{def} \Box_A\langle A\rangle\eta$.

Proof. We show that for this fragment, the valuation of formulas becomes 'moment determinate', that is, for all strategy / state pairs with the same state (moment), they evaluate to the same truth value (see Horty [6] for further explanation of this terminology). First note that the truth condition for the combined ('fused', as Horty calls it) operator $\Diamond_A[A]\eta$, reduces to the following moment determinate truth condition.

$\mathcal{M}, \alpha_A, s \models \Diamond_A[A]\phi U^{ee}\psi \Leftrightarrow \exists \beta_A$ such that $\forall \gamma_{\overline{A}}$ it holds that $\exists n > 0$ such that
(1) $\mathcal{M}, \alpha_A, (F_{\beta_A | \gamma_{\overline{A}}})^n(s) \models \psi$ and
(2) $\forall i$ with $0 < i < n$ we have $\mathcal{M}, \alpha_A, (F_{\beta_A | \gamma_{\overline{A}}})^i(s) \models \varphi$

This truth condition is completely independent of the strategy α_A. For similar reasons the truth condition for the combined operator $\Box_A\langle A\rangle\eta$ is moment determinate. Now notice that also all other formulas of the sub-language determined by $\langle\langle A\rangle\rangle\eta \equiv_{def} \Diamond_A[A]\eta$ and $[[A]]\eta \equiv_{def} \Box_A\langle A\rangle\eta$ are moment determinate. This means the quantification over all strategy / state pairs in the definition of validity gives the same result when performed only with respect to all states (moments). It is not too difficult to see that we thus arrive at a concise, but correct semantics for ATL.

Proposition 1. *The logic of the operators $\Box_A\varphi$ is S5 for every set A.*

This is due to the fact that S5 is sound and complete for equivalence classes. The accessability relation for the modal operator $\Box_A$ is the relation connection alternative A strategies. For any given model the 'alternative relation' forms a fixed equivalence class. As a consequence we have validities such as

$$\models [A]\eta \rightarrow \Diamond_A[A]\eta$$

saying that if agents A strategically see to it that η, indeed they have the ability to do so, and

$$\models \Box_A\langle A\rangle\eta \rightarrow \langle A\rangle\eta$$

saying that if for all strategies it is the case that agents A may encounter η, they currently perform a strategy where they possibly encounter η. It also follows that nesting of operators $\Box_A$ and $\Diamond_A$ is not meaningful, since it is well-known that nested S5 formulas can be replaced by logically equivalent non-nested formulas.

Proposition 2. *The operators $\Box_A\varphi$ obey the interaction axioms:*

$$\models \Box_A\varphi \rightarrow \Box_B\Box_A\varphi$$

$$\models \Diamond_A\varphi \rightarrow \Box_B\Diamond_A\varphi$$

Below we list only a few more validities. Possible complete axiomatizations for the present logic are still under investigation.

Proposition 3. *Additionally, we have the following validities and non-validities.*

$$\models [A]\eta \rightarrow [B]\eta \text{ for } A \subseteq B$$
$$\models \langle A\rangle\eta \rightarrow \langle B\rangle\eta \text{ for } A \supseteq B$$
$$\models [A]X\varphi \wedge [B]X\psi \rightarrow [A \cup B]X(\varphi \wedge \psi)$$
$$\models \langle A \cup B\rangle X(\varphi \wedge \psi) \rightarrow \langle A\rangle X\varphi \vee \langle B\rangle X\psi$$

Note that for the third validity, we do not need the condition of sets A and B being disjoint, as in the axiomatizations of CL [12] and ATL.

4 Epistemic ATL-STIT

As a demonstration of the applicability of the formalism, we extend it with epistemic modalities. We interpret the epistemic modalities using epistemic indistinguishability relations over over strategy / state pairs. The resulting fine-grained epistemic structures enable us to shed new light on the problem of so called 'uniform strategies'.

4.1 Basic Definitions

First we extend the language of ATL-STIT with an operator $K_a\varphi$ for agent a knows φ, an operator $E_A\varphi$ for agents A all know that φ, an operator $D_A\varphi$ for agents A would know that φ if they would exchange all their knowledge, and an operator $C_A\varphi$ for agents A commonly know that φ.

Definition 8. *Well-formed formulas of the temporal language $\mathcal{L}_{E\text{-}ATL\text{-}STIT}$ are defined by:*

$$\varphi, \psi, \ldots := p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid E_A\varphi \mid D_A\varphi \mid C_A\varphi \mid \Diamond_A\varphi \mid \Box_A\varphi \mid [A]\eta \mid \langle A\rangle\eta$$
$$\eta, \theta, \ldots := \phi U^{ee} \psi$$

To accommodate epistemic reasoning, we want to define S5 indistinguishability relations over the units of evaluation, that is, strategy / state pairs. However, we have to be careful. As pointed out before, in for instance [13], adding epistemic indistinguishability relations to arbitrary ATSs leaves room for ambiguity: in particular, what is the epistemic status of an action leading from one state to another one that is epistemically indistinguishable? Should we interpret this as the agents not being able to recall the action? Or do they recall the action, but only do not know the resulting and originating state? To avoid this ambiguity, we can better add epistemic relations to ATSs that are trees.

Definition 9. *An ATS $\mathcal{M} = (S, \mathcal{T}, \pi)$ is an ATS where the function $\mathcal{T}$ is such that the system relation $\mathcal{R}^{sys}$ is a tree.*

Now note that on the subclass of tree-ATSs, the definitions of section 3.2 result in exactly the same logic ATL-STIT. This is because any ordinary ATS can be unravelled into a tree-ATS that is modally indistinguishable.

Now we can add the epistemic indistinguishability relations for separate agents. This results in a most general setup for the semantics of E-ATL-STIT, where beforehand nothing is determined about whether agents recall their actions or not: if there is an epistemic indistinguishability relation between two subsequent states of a fixed strategy, the agents cannot recall having done that action, but if there is not such a relation, they can.

Definition 10. *We extend models $\mathcal{M} = (S, \mathcal{T}, \pi)$ to models $\mathcal{M} = (S, \mathcal{R}_A, \mathcal{T}, \pi)$. The relation $\mathcal{R}_a$ for individual agents a is an equivalence relation over strategy / state pairs (α_A, s).*

We can define any of the multi-agent versions of knowledge, that is, distributed (or implicit) knowledge, shared knowledge (everybody knows) and common knowledge (reflexive transitive closure of shared knowledge), in terms of the indistinguishability relations over strategy / state pairs for the individual agents. In the standard way, we extend the truth definitions with the following clauses for the (group) knowledge operators.

Definition 11.

$$\mathcal{M}, \alpha_A, s \models K_a\varphi \Leftrightarrow \forall(\beta_B, t) \text{ with } (\alpha_A, s)\mathcal{R}_a(\beta_B, t) \text{ it holds that } \mathcal{M}, \beta_B, t \models \varphi$$
$$\mathcal{M}, \alpha_A, s \models E_A\varphi \Leftrightarrow \forall(\beta_B, t) \text{ with } (\alpha_A, s)(\bigcup_{a\in A} \mathcal{R}_a)(\beta_B, t) \text{ it holds that } \mathcal{M}, \beta_B, t \models \varphi$$
$$\mathcal{M}, \alpha_A, s \models D_A\varphi \Leftrightarrow \forall(\beta_B, t) \text{ with } (\alpha_A, s)(\bigcap_{a\in A} \mathcal{R}_a)(\beta_B, t) \text{ it holds that } \mathcal{M}, \beta_B, t \models \varphi$$
$$\mathcal{M}, \alpha_A, s \models C_A\varphi \Leftrightarrow \forall(\beta_B, t) \text{ with } (\alpha_A, s)((\bigcup_{a\in A} \mathcal{R}_a)^*)(\beta_B, t) \text{ it holds that } \mathcal{M}, \beta_B, t \models \varphi$$

The above proposal for adding the epistemic dimension is very general. Clearly it results in an S5 logic for individual agent knowledge, while leaving the sub-logic of ATL-STIT in tact. Of course several intuitive extra relational properties can be considered, leading to specific interaction properties. However, for our discussion on uniform strategies, below, the definitions suffice.

4.2 The Problem of Uniform Strategies

The most discussed problem for epistemic additions to ATL discussed in the literature (ATEL [14]), is the problem of so called 'uniform strategies'. We briefly recall the problem by means of the cards example from [13] (which we slightly adapt). There is a deck of three cards, A, K and Q. There is a somewhat unconventional order on these cards, where A beats K, K beats Q, but Q beats A. Now consider two gambling agents a and b who each get a card from the dealer. Before a showdown occurs, agent a is given the choice to swap his card with the one remaining on the dealers deck. Apparently due to the incompleteness of his knowledge a does not know a winning strategy. He does not know the card still in the deck, but depending on what this card is, he either has to swap or not in order to win. Structures of ATEL equip ATSs with epistemic indistinguishability relations between states (moments). Now it is perceived as counterintuitive that in the ATEL structures we can draw for this little game, at the moment corresponding to the decision point of agent a, it is true that $K_a\langle\langle a\rangle\rangle win$. This holds since the agent cannot distinguish the state where he has the winning card from the state where he has the loosing card, but whichever state he is in, it has a guaranteed possibility to win if it chooses the right strategy in the right state. However, the truth of this formula is perceived as counterintuitive since one is tempted to believe that it expresses that a has a *single* 'uniform strategy' for winning, that is, a strategy that guarantees a win irrespective of the state the agent is in.

But it appears to us that if we stay faithful to the intended meaning of the operators involved, the formula is not counterintuitive: it exactly expresses what is the case, namely that agent a knows that there is a strategy to win. Indeed that does not imply that he knows what strategy to apply, which, in this case, is exactly the only reason why he cannot ensure the win. So, the problem appears to be that one is tempted to read something in the formula that is not there, namely, that the agent knows a uniform strategy for winning. Maybe the present formalism, that decomposes the standard ATL operators in two separate modal operators, enables us to see that more clearly.

However, an ensuing problem is that one indeed would like to have a way of expressing that an agent, or group of agents does not know what the current state is, while at the same time they do know (or do not know) how to win. In the above example, the agent a did not know how to win. We would like to have a formula corresponding to that fact. In ATEL [14] we cannot express that. But the present formalism, with its more fine grained epistemic structures, enables us to express this directly as $\neg\Diamond_a K_a[a]win$, that is, a has no single strategy for which he knows he is guaranteed to win. We cannot find an equivalent formula in ATEL, because ATEL's semantic structures are not fine-grained enough in two respects. First, because in ATEL, evaluation is only with respect to states, it cannot give semantics to the decomposition of the ATL operator $\langle\langle A\rangle\rangle\eta$ into $\Diamond_A[A]\eta$, and second, because epistemic indistinguishability relations are

defined over states, it cannot give semantics to the notion of an agent knowing a strategy.

Then the question is, does this solve the problem of so called 'uniform strategies' as formulated in the literature? That depends on how one looks at it. Actually it is not completely clear to us what in the context of ATSs, should be understood by a 'uniform strategy'. The notion of 'uniform strategy' comes from game theory [15]. But game theory is different from logic in that it studies the properties of game structures as such, that is, independent of a logical language like ATL to be interpreted over them. In game structures the choices have action names. ATL, and also STIT-ATL are endogenous temporal formalisms that cannot express anything related to the action names of game structures. And in particular those action names have been associated to the notion of 'uniform strategies'. Uniform strategies have been described as strategies where the 'same actions' are performed from different states to ensure a certain property. If actions have names, the same actions can be defined as actions having corresponding names. The present proposal does not solve the problem of uniform strategies interpreted in this sense. We believe, solutions would require an exogenous language, where in one way or the other there is reference to the names of actions in the object language. However, in a weaker sense the present proposal does solve the problem. In ATSs actions are identified with what they bring about. Then, typically, single strategies take *different* actions from different states. And it is also the other way around: taking two different strategies in two different states may mean that one performs the same actions. Now, if 'knowing a uniform strategy for φ, without possibly knowing the current state' is defined as 'knowingly seeing to it that φ, without possibly knowing the current state', the present proposal does offer a solution to the problem of uniform strategies.

Generalizing the idea in [7] we can express that there is an A-strategy, where the agents A commonly know that they ensure η as:

$$\Diamond_A C_A[A]\eta$$

Agents A commonly knowing the existence of a strategy (without knowing whether they actually perform the strategy) is expressed as:

$$C_A \Diamond_A[A]\eta$$

Note that in the first of the above formulas, for the concept of 'a group of agents A knowingly performing a strategy', we used that the agents have *common knowledge* that they perform the strategy. We thus defined this concept as $C_A[A]\eta$. In our opinion distributed knowledge or shared knowledge is not enough. For instance, me and a friend can only knowingly follow a strategy of meeting in Paris someday if I know that he knows, and I know that he knows that I know, etc.

5 Related Research

Horty ([6] p. 151) explains that it is not that easy to generalize the standard STIT framework where evaluation is with respect to moment / history pairs, to the strategic case.

In general, more than one strategy may be compatible with the same moment / history pair. Horty's first suggestion is then to implicitly quantify over all strategies that correspond to a given moment / history pair. His second suggestion is much closer to the solution proposed in this paper (note that here we assume the close relatedness between the STIT-framework and the ATL-framework we explored in [10]). Horty suggests to evaluate formulas with respect to 'state / history / history-set' triples (where the history is an element of the history-set), and to define the semantics of his strategic STIT operator $[A\ cstit : \varphi]$ (agents A strategically see to it that φ) as there being a strategy α, such that the history-set equals the histories admitted by the strategy, and φ being true on all these histories. Our proposal differs from this proposal on three points. First, for the present ATL-setting we do not see the need to include the history in the units for evaluation. Second, we think it is better to simply see the strategies themselves as part of the units of evaluation. We explicitly need this in our discussion of uniform strategies in section 4.2. Finally, we believe Horty's definition fails to model the important property that if a set of agents sees to something, any superset also sees to that same something. This property follows from our definition as the result of taking the intersections in the truth conditions for $[A]\varphi$ and $\langle A\rangle\varphi$.

Using ideas similar to ours Müller [8] defines a semantics for the notion of 'continuous action' in the STIT framework. Like us, Müller suggests to take up strategies as elements in the units over which to evaluate formulas. To be more precise, Müller evaluates with respect to 'context-state / state / history / strategy' quadruples. His notion of ISTIT (*is* seeing to it that), is then defined, roughly, as truth on all histories admitted by the strategy. Although the idea to take up strategies in the units of evaluation is similar, other aspects of the approach are quite different. That is not too surprising, since Müller's aim is an ISTIT operator, while we aim at a strategic STIT operator. Also Müller does not aim at defining a multi-agent variant of his operator. More in particular, his strategies are always single agent strategies. In our setting, the problem of dealing with multi-agent strategies is central.

Finally, also Jamroga and gnotes [9] suggest to change the units of evaluation. Aiming at solving the problem of uniform strategies in ATEL, they suggest to evaluate formulas with respect to sets of states. However, their approach is much further removed from our approach than Horty's or Müller's.

6 Conclusion

This paper extends ATL with strategic STIT operators. We argued that the evaluation with respect to strategy / state pairs is essential for a logic that aims to reason about decisions that are fixed for groups of agents. Here the decisions are to take a particular strategy. Also we discussed the problem of uniform strategies, and explained how our formalism can be seen as a partial answer to that problem.

There are many possible applications of this extended formalism. We discussed some preliminary investigations in the epistemic realm. Another route of investigation is the extension with deontic operators. One of the reasons STIT logics are popular in deontic logic is that they are the best formalism around to model the fourth sentence of Chisholm's infamous benchmark scenario for deontic formalizations [16]. To add

deontic expressivity, we may consider several options. For instance, Wansing [17] has suggested to model personal obligations imposed by one agent onto the other by identifying this with 'agent *a* sees to it that agent *b* is punished if he does not comply to his obligations'. This approach can incorporated in the present framework very well. Another option is simply to define a deontic accessibility relation over strategy / state pairs, like we did for the epistemic indistinguishability relation.

References

1. Alur, R., Henzinger, T., Kupferman, O.: Alternating-time temporal logic. In: FOCS '97: Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS '97), IEEE Computer Society (1997) 100–109
2. Alur, R., Henzinger, T., Kupferman, O.: Alternating-time temporal logic. Journal of the ACM **49**(5) (2002) 672–713
3. Belnap, N., Perloff, M.: Seeing to it that: A canonical form for agentives. Theoria **54** (1988) 175–199
4. Belnap, N., Perloff, M.: Seeing to it that: A canonical form for agentives. In Kyburg, H.E., Loui, R.P., Carlson, G.N., eds.: Knowledge Representation and Defeasible Reasoning. Kluwer, Boston (1990) 167–190
5. Belnap, N., Perloff, M., Xu, M.: Facing the future: agents and choices in our indeterminist world. Oxford University Press (2001)
6. Horty, J.: Agency and Deontic Logic. Oxford University Press (2001)
7. Herzig, A., Troquard, N.: Knowing How to Play: Uniform Choices in Logics of Agency. In Weiss, G., Stone, P., eds.: 5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan, ACM Press (2006) 209–216
8. Müller, T.: On the formal structure of continuous action. In Schmidt, R., Pratt-Hartmann, I., Reynolds, M., Wansing, H., eds.: Advances in Modal Logic. Volume 5., King's College Publications (2005) 191–209
9. Jamroga, W., gotnes, T.: Constructive knowledge: what agents can achieve under incomplete information. Technical Report IfI-05-10, Institute of Computer Science, Clausthal University of Technology, Clausthal-Zellerfeld (2005)
10. Broersen, J., Herzig, A., Troquard, N.: From coalition logic to stit. In: Proceedings LCMAS 2005. Electronic Notes in Theoretical Computer Science, Elsevier (2005)
11. Goranko, V., Jamroga, W.: Comparing semantics of logics for multi-agent systems. Synthese **139**(2) (2004) 241–280
12. Pauly, M.: A modal logic for coalitional power in games. Journal of Logic and Computation **12**(1) (2002) 149–166
13. Jamroga, W., Hoek, W.v.d.: Agents that know how to play. Fundamenta Informaticae **63**(2) (2004)
14. Hoek, W.v.d., Wooldridge, M.: Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. Studia Logica **75**(1) (2003) 125–157
15. Neumann, J.v., Morgenstern, O.: Theory of games and economic behaviour. Princeton University Press (1944)
16. Chisholm, R.: Contrary-to-duty imperatives and deontic logic. Analysis **24** (1963) 33–36
17. Wansing, H.: Obligations, authorities, and history dependence. In Wansing, H., ed.: Essays on Non-classical Logic. World Scientific (2001) 247–258