

A Reading Companion to the ESSLLI Course "Logics for Agency and Multi-Agent Systems"

Jan Broersen Andreas Herzig Nicolas Troquard

Abstract

This paper is a reading companion to our introductory ESSLLI course on "Logics for Agency and Multi-Agent Systems". Proofs are omitted. This reader's aim is mostly to structure the material that will be presented at the course, and to point the attendees of the course to the relevant literature.

Contents

1	Modal logic for Multi-Agent Systems	3
1.1	Recap of basic logic notions	3
1.2	Normal Modal Logic: the very basics	4
1.2.1	The (multi-)modal language	4
1.2.2	Normal modal logics, axiomatically	4
1.2.3	The modal semantics	5
1.2.4	Modal logic consequence	6
1.3	Expressivity	7
1.3.1	On the level of models: bisimulation invariance	7
1.3.2	On the level of frames: correspondence theory	7
1.4	Soundness and completeness	8
1.5	Example of combining modal logics: products	10
1.6	Decidability and complexity issues	10
1.6.1	Decidability	10
1.6.2	Complexity	11
1.7	Non-normal modal logics: neighborhood models	11
1.8	Modal logic: application to some reasoning domains	12
1.8.1	Dynamic Logic	12
1.8.2	Epistemic Logic	12
1.8.3	STIT logic	13
1.8.4	Deontic Logic	13
1.8.5	Linear Time Temporal Logic	13
2	Logical Frameworks for Multi-Agent Systems: Cohen & Levesque	15
2.1	Motivation and background	15
2.2	Syntax	15
2.3	Semantics	15
2.3.1	The ingredients	15
2.3.2	The models	16
2.4	Achievement goal, persistent goal and intention	17
2.5	Discussion of C&L	17

3	The power of cooperation: Coalition Logic	19
3.1	Motivation	19
3.2	Syntax	19
3.3	Semantics: neighborhood models	19
3.4	Semantics: game structures	20
3.5	Axiomatization	20
3.6	Discussion: epistemic extensions and their problems	21
4	STIT theory of agency and applications	23
4.1	Motivation	23
4.2	STIT models: BT+AC	23
4.3	Semantical comparison of CL and STIT	25
4.3.1	Initial settings	25
4.3.2	Translating models: By the example	26
4.3.3	A translation from CL to discrete STIT	26
4.4	Mathematics of STIT	27
4.4.1	Axiomatizing individual Chellas’s STIT (CSTIT)	27
4.4.2	Extension to groups of agents (\mathcal{G} STIT)	28
4.5	STIT embraces CL in the realm of normal modal logics	29
4.5.1	Normal Simulation of Coalition Logic (NCL)	30
4.5.2	From CL to NCL	30
4.6	Introducing uncertainty	30
4.6.1	The Conformant STIT (ENCL)	31
4.6.2	Reasoning about uniform choices	31
5	Intention revisited: enhancing Cohen and Levesque	33
5.1	A logic of agency and mental states	33
5.2	Intention to be	34
6	Going fully strategic	35
6.1	Strategic ability: Alternating Time Temporal Logic	35
6.1.1	Syntax, semantics and axiomatization of ATL	35
6.1.2	Game structures vs. alternating transition systems	37
6.1.3	Coalition Logic and CTL as fragments of ATL	37
6.2	Embedding ATL into strategic STIT ability	38
6.3	Strategic STIT	41
6.3.1	Core Syntax, Abbreviations and Intended Meanings	41
6.3.2	Model theoretic semantics	43
6.4	Epistemic strategic STIT	45
6.4.1	Basic definitions	45
6.4.2	The problem of uniform strategies	46

1 Modal logic for Multi-Agent Systems

There are many definitions for what a multi-agent system is [Woo02]. For the purpose of this course, we define a multi-agent system as a set of acting and interacting, deliberating and communicating, autonomous, goal directed and socially engaged computing components. The idea that logic is a valuable tool for understanding, describing and, ultimately, programming multi-agent systems, is as old as the paradigm of multi-agent systems itself. In the first part of this ‘reading companion’ to the ESSLLI course "Logics for Agency and Multi-Agent systems" we give a brief overview of important logical frameworks that have been developed in this area. However, we do not want to do that without first having recalled some of the basic techniques of logic.

In the second part, starting with section 4, we discuss one of the most prominent philosophical viewpoints on agency, called STIT theory. We discuss the incorporation of this view into logical frameworks for multi-agent systems. The interest of the computer science community working on multi-agent systems into the philosophical work on logics for agency has emerged only recently.

We will confine ourselves to modal logics in this course. This is because these logics are most widely used for describing multi-agent systems. The question of what modal logic is can be answered in many different ways. But before answering this question, we first want to recall briefly what a logic is.

1.1 Recap of basic logic notions

Logics are formalizations of possible reasoning patterns. Different logics apply to different reasoning contexts. And, as we see it, formalization is a prerequisite for the philosophical discussions on which logics best fit which reasoning domains.

The definition of any formal logic starts with the definition of a formal language. For propositional logic (PL) this is, in BNF notation, with p ranging over a countably infinite set ATM of proposition symbols:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi$$

Other propositional connectives, such as the material implication, are defined as abbreviations: $\varphi \rightarrow \psi \equiv_{def} \neg\varphi \vee \psi$. The logic itself is then defined as a particular subset of formulas in this language. These formulas are the logical invariants of the reasoning domain the logic is designed for. This subset of logical invariants can be described in at least two fundamentally different ways.

The first one is the axiomatic approach. This works with a set of fixed axiom schemas and a set of rules deriving theorems from them. Together, the axiom schemas and rules determine the subset of formulas of the language that form the logic. We explicitly talk about axiom schemas here. Schemas allow us to describe infinite sets of formulas by finite means. Of course most logics, as subsets of infinite languages, are infinite themselves. If we want to capture them by finite means, we can use schemas. For instance, with the schema $\varphi \vee \neg\varphi$ of propositional logic we generate an infinite number of axioms by uniformly substituting formulas of the language for φ . Often, axiom schemas are also simply referred to as ‘axioms’. In the sequel, we will also sometimes do that.

Important logical notions for the axiomatic view on logic are (1) theoremhood, and (2) consistency. A formula φ is a theorem if it can be derived from the axioms Γ using the inference rules of the logic, notation $\Gamma \vdash_L \varphi$. Consistency of a formula φ is defined as $\Gamma \not\vdash_L \neg\varphi$. In this case we talk of ‘ L -consistency’ of the formula. Consistency of the *logic* L is defined as consistency for all its theorems. It follows that in a consistent logic, there is never a formula φ such that $\Gamma \vdash_L \varphi$ and $\Gamma \vdash_L \neg\varphi$.

The second, semantic approach to defining a logic is based on model theory. Important notions from the semantic view on logic are (1) validity, and (2) satisfiability. Starting with the notion of a ‘model’, and the definition of what it is for a formula to be *true* on a model, we arrive at the definition of *valid* formulas as those who are true on any model. Recall that a propositional model is a function $V : ATM \rightarrow True, False$ assigning to each atomic proposition either the value *True* or the value *False*. A propositional formula is true on a model if it evaluates to true according to the truth functional semantics of the propositional operators. A formula is propositionally valid, also called ‘a

tautology’, if it is true on all possible models. The notion of ‘satisfiability’ is the semantic counterpart of consistency. Indeed if a logic is sound and complete, theorems derivable using the axiomatization correspond to validities of the semantic description, and consistent formulas correspond to satisfiable formulas.

Now we have to come back to our statement about what a logic is. We said that logics are formalizations of possible reasoning patterns. However, it is not clear beforehand how a logic, defined as a subset of a certain language, can be seen as a formalization of the reasoning patterns of some domain. To see this, we need to introduce the notion of logical consequence. Again we have two views on this notion; a semantic and a syntactic one.

In propositional logic, the semantic notion of logical consequence is defined as follows. φ is a logical consequence of the set of formulas Ψ , notation $\Psi \models_{PL} \varphi$, if all propositional models that satisfy all formulas in the set Ψ are also models for φ . The corresponding syntactic variant of logical consequence is simply to add the specific formulas Ψ as axioms to the schemas of the logic, and declare φ a consequence, notation $\Psi \vdash_{PL} \varphi$ if it can be derived as a theorem from this extended axiom system. For modal logic this cannot be straightforwardly generalized, as we will see in the next section.

The task of *checking* logical consequence can often be reduced to the task of checking satisfiability. That is, checking if $\Psi \models \varphi$ is often the same as checking if $(\bigwedge \Psi) \wedge \neg\varphi$ is *not* satisfiable. This holds, for instance, for PL. But, it does not work for all possible notions of logical consequence. In particular, it holds only for consequence notions for which we have the deduction theorem: $\Phi \models \psi$ iff $\models (\bigwedge \Phi) \rightarrow \psi$, which indeed holds for propositional logic and finite sets Φ .

1.2 Normal Modal Logic: the very basics

There are many possible answers to the question of what modal logic is. Even to the question what the language of modal logic is, different answers are possible. For this course it suffices to consider the multi-modal languages.

1.2.1 The (multi-)modal language

The language of basic modal logic is the language of propositional logic extended with the unary operators \Box and \Diamond . But, often it is possible to use the abbreviation $\Diamond\varphi \equiv_{def} \neg\Box\neg\varphi$, by which the box can be defined in terms of the diamond or the other way round. The language of multi-modal logic provides an infinite set of such modal boxes. Formally, in BNF notation, with i ranging over a infinite set *Labels*:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \Box_i\varphi$$

Diamonds we define through abbreviation: $\Diamond_i\varphi \equiv_{def} \neg\Box_i\neg\varphi$.

1.2.2 Normal modal logics, axiomatically

Now the syntactic answer to what a *normal* modal logic is, is as follows. Any axiom system that contains the schemas of *PL*, the schema:

$$(K) \quad \Box_i(\varphi \rightarrow \psi) \rightarrow (\Box_i\varphi \rightarrow \Box_i\psi)$$

and the rules:

(**Modus Ponens**) from φ and $\varphi \rightarrow \psi$ infer ψ

(**Necessitation**) for every i , from φ infer $\Box_i\varphi$

is a normal multi-modal logic (note that since the axioms are *schemas*, we may instantiate the ‘variables’ φ and ψ by arbitrary formulas, as long as we do this uniformly for a schema). The basic minimal modal system determined by the above axiomatization is referred to as the system K. The K refers to the Kripke models used for the semantics.

1.2.3 The modal semantics

The semantics of normal multi-modal logics is given in terms of Kripke models. Kripke models themselves are based on Kripke frames.

A Kripke frame F for the multi-modal language is a triple $F = \langle W, R \rangle$, where

W is a non-empty set of possible worlds,

R is a function $R : Labels \rightarrow (W \times W)$ mapping each modal index to a binary relation over worlds

Based in a Kripke frame F we can make Kripke models.

Given a finite set of proposition symbols ATM , a frame $F = \langle W, R \rangle$ is extended to a model $M = \langle W, R, V \rangle$ where:

V is a function $V : W \rightarrow 2^{ATM}$ mapping each world to a subset of atomic proposition symbols.

Strictly speaking, we do not need the intermediate notion of ‘frame’ to give semantics to modal languages. But it turns out that properties of modal logics stronger than the basic modal logic are intimately related to the properties of frames. This is the subject of ‘correspondence theory’ that we discuss in section 1.3.2.

Kripke models, with their relations over possible worlds, turn out to be extremely valuable semantic structures that can function as abstract representations for many interesting reasoning domains. For instance, the structure of time is easily thought of as a set of moments related by time itself. Or, we can take the notion of a possible world directly to mean that a world is considered possible according to an agent’s knowledge. Any relation between two worlds then represents whether from the viewpoint of one world, the other world is an epistemic possibility. In applications of modal logic like these, that we will discuss in a little more detail in section 1.8, the box operator always expresses some form of ‘necessity’, while the diamond operator expresses ‘possibility’. The formal counterpart of this generalized interpretation of the box operator is expressed by the following truth condition (note that we use infix notation for relations).

Given a model $M = \langle W, R, V \rangle$, a modal formula of the form $\Box_i \varphi$ is said to be *true* with respect to a world w in W , notation $M, w \models \Box_i \varphi$, if for all $w' \in W : wR(i)w'$ implies that $M, w' \models \varphi$.

Although we introduced diamonds by abbreviation, it is useful to also take note of the truth condition for diamonds that we could have used would we not have introduced it in that way.

Given a model $M = \langle W, R, V \rangle$, a modal formula of the form $\Diamond_i \varphi$ is said to be *true* with respect to a world w in W , notation $M, w \models \Diamond_i \varphi$, if there is a $w' \in W$ such that $wR(i)w'$ and $M, w' \models \varphi$.

We assume the truth functional interpretation for propositional operators to be known. To define what a normal modal logic is semantically, we have to take three more steps. At this point we have only defined truth in a world of a model. We generalize this notion now in three steps to arrive at general validity of a formula, which corresponds to membership of the logic.

A formula φ is *valid on a Kripke model* M , notation $M \models \varphi$, if for all w in W it holds that $M, w \models \varphi$.

A formula φ is *valid on a Kripke frame* F , notation $F \models \varphi$ if for all M based on F it holds that $M \models \varphi$.

A formula φ is *generally valid*, notation $\models \varphi$ if for all F of a class of frames \mathcal{F} , we have that $F \models \varphi$.

Now we can define normal modal logics as those subsets of the modal language that consist of formulas which are generally valid with respect to some class \mathcal{F} of Kripke frames. The defining properties of the class of frames give rise to modal validities. The languages in which the properties of frame classes are defined, are usually first- or second-order logic. The logic of the most general class of Kripke frames, is the logic K. But if we narrow this class to a class with certain first-order second-order definable properties, additional formulas may be valid, resulting in stronger logics. Some well known first-order properties often appearing in modal semantics are:

$F = \langle W, R \rangle$ is reflexive if $\forall w \in W : wRw$.

$F = \langle W, R \rangle$ is transitive if $\forall s, t, u \in W : sRt \wedge tRu \rightarrow sRu$.

$F = \langle W, R \rangle$ is symmetrical if $\forall s, t \in W : sRt \rightarrow tRs$.

$F = \langle W, R \rangle$ is Euclidean if $\forall s, t, u \in W : sRt \wedge sRu \rightarrow tRu$.

$F = \langle W, R \rangle$ is serial if $\forall s \in S, \exists t \in W : sRt$.

$F = \langle W, R \rangle$ is an equivalence relation if R is reflexive, transitive and symmetrical

The classes of frames described by certain first-order conditions are given special names in modal logic. T is the class of all reflexive Kripke models. S4 is the class of all reflexive-transitive Kripke models. S5 is the class of all Kripke models with accessibility relations that are equivalence relations. KD45 is the class of all Kripke models with serial, transitive and Euclidean accessibility relations. The modal logics generated by these frames have similar names: KT, S4, S5, KD45, etc.

Whether we can describe some logic both axiomatically and semantically as the logic over some particular class of Kripke frames, is the concern of soundness and completeness results, that we discuss in section 1.4.

1.2.4 Modal logic consequence

If we want to generalize the standard notion of entailment (logical consequence) given in 1.1 to modal logic, we have several options. But the most widely used notion of modal logic consequence is the following.

φ is a modal logic consequence of a set of formulas Ψ , notation $\Psi \models \varphi$, if for all Kripke models M and all possible worlds w , it holds that if $M, w \models \psi$ for all $\psi \in \Psi$, then $M, w \models \varphi$.

Since the models in the above definition are defined relative to a class of frames \mathcal{F} , whether or not a modal formula is derivable from a set of formulas depends on the properties of the frames of the logic.

The corresponding syntactic operation is not straightforward. The definition for propositional logic cannot be generalized to the modal case without adaptation. The problem is that the rule of necessitation is not compatible with the above semantic ‘local’ version of entailment. In particular, we do not want to derive from a fact p that $\Box p$, because, from the observation that it rains, we do not want to derive that it rains necessarily. Apparently, we only want to use the rule of necessitation on theorems of the logic, and not on facts about the reasoning domain. A solution is to constrain derivations in the following way: in proofs that derive logical consequences of sets of facts, never use an axiom corresponding to a plain fact of the domain before using a necessitation rule. This clearly eliminates the above problem. And the resulting syntactic notion of derivability corresponds with the semantic, local notion of entailment.

There is consequence relation, namely the *global* one defined in terms of validity on models, that does correspond to the unlimited use of necessitation. However, we will not use this stronger notion of consequence. Also, for the above defined notion of *local* consequence we have the deduction theorem: $\Phi \models \psi$ if and only if $\models (\bigwedge \Phi) \rightarrow \psi$. For stronger notions of consequence, we do not have this.

1.3 Expressivity

Before we discuss soundness and completeness in the next section, we need to explain some issues concerning the expressiveness of modal logic. We consider expressivity with respect to models and with respect to frames. On the level of models we are concerned with the discriminative power of *individual* modal logics; what modal *formulas* distinguish between what *models*? On the level of frames we are interested in the discriminative power of modal logics as a *family* of logics; what modal *schemas* correspond to what properties of *frames*?

1.3.1 On the level of models: bisimulation invariance

Bisimilarity is a relation between worlds or between models. It captures the idea of ‘semantic equivalence’. Bisimilar worlds are semantically equivalent for modal logic in the sense that no modal formulas exist that ‘distinguish’ between bisimilar worlds. Semantical indistinguishability means that if world w of model M is bisimilar to world w' in model M' , we cannot find a modal formula that is true in M, w but false in M', w' .

A world w^1 of model M^1 is *bisimilar* to a world w^2 in model M^2 if and only if:

- (1) $V^1(w^1) = V^2(w^2)$,
- (2) $\forall w'^1 \in W^1 : w^1 R^1(i)w'^1$ implies $\exists w'^2 \in W^2 : w^2 R^2(i)w'^2$ and w'^1 and w'^2 are bisimilar,
- (3) $\forall w'^2 \in W^2 : w^2 R^2(i)w'^2$ implies $\exists w'^1 \in W^1 : w^1 R^1(i)w'^1$ and w'^2 and w'^1 are bisimilar.

The notion of bisimilarity can be used to show that certain properties of models are not expressible. An example is irreflexivity ($\forall w \in W, \neg wRw$) of models. We cannot *characterize* this property at the level of *models*, because for any point in a model that is reflexive, we can always construct a bisimilar point by unraveling.

The notion of bisimilarity goes to the heart of what modal logic is. Van Benthem actually *defined* modal logic to be the fragment of first order logic that is invariant under bisimulation. To understand this characterization, we would have to explain how modal logic can be seen as a fragment of first-order logic. We leave it to the reader to look this up in for instance [BRV01].

Note that this does not mean that worlds for which any possible modal formula evaluates to the same value, are bisimilar. Indeed one can construct infinite models for which this is not the case. Again, the interested reader is referred to [BRV01].

1.3.2 On the level of frames: correspondence theory

Correspondence theory concerns two questions. The first is: which frame class generates a given set of validities (usually given as a schema)? In other words, if we are given a set of validities, can we find a first- or second order property C of frames such that the validities are in the modal logic of this frame? The second question is the one we already encountered in section 1.2.3: which modal validities are generated by a certain frame class? In other words, which validities hold on all frames with first-, or second order property C ? If we have *both* ways, then we say the schema *corresponds* to, or *characterizes* the property C on frames. Some example of well-known correspondences are the following.

Example1: the schema $\Box_i \varphi \rightarrow \Box_i \Box_i \varphi$ *corresponds* to the class of transitive frames

Example2: the schema $\Box_i \varphi \rightarrow \varphi$ *corresponds* to the class of reflexive frames

Example3: the schema $\varphi \rightarrow \Box_i \Diamond_i \varphi$ *corresponds* to the class of symmetric frames

Note that these formulas say little about similar correspondences with classes of *models*: it is not too difficult to find a model that obeys $\Box_i \varphi \rightarrow \varphi$, and that is not reflexive. Take for instance an infinite series of worlds with the same valuations of atoms related through $R(i)$ in a linear order. All worlds in this series are bisimilar, and thus satisfy $\Box_i \varphi \rightarrow \varphi$.

Not all first-order properties of frames correspond to a modal validity schema. An example is again irreflexivity. It certainly does not correspond to the schema $\varphi \rightarrow \Box_i \neg \varphi$. Although this schema *seems* to express one side of correspondence, that is, that if it is true on a frame, the frame cannot be reflexive, it is actually an example of a schema that cannot be valid on *any* frame. Indeed, not all schemas are expressible as properties of frames. Also, among the schemas that *are* expressible as properties of frames, there are some that are not *first-order* expressible. Examples are (1) Löb's axiom $\Box_i(\Box_i \varphi \rightarrow \varphi) \rightarrow \Box_i \varphi$ that corresponds with ' $R(i)$ is transitive and conversely well-founded', and (2) Mc Kinsey's axiom $\Box_i \Diamond_i \varphi \rightarrow \Diamond_i \Box_i \varphi$. Both these schemas are *only* expressible as second-order properties of frames.

1.4 Soundness and completeness

The issue of soundness and completeness of modal logics concerns the question whether semantic characterizations and axiomatic descriptions are identical. In particular

Soundness: if $\vdash \varphi$ then $\models \varphi$

Using induction over the axiomatic system, soundness is usually fairly easy to prove. First one has to prove that axioms are validities, and second that rules preserve validity. In the early days, modal logic was an entirely axiomatic enterprise. But there were no means for proving that the systems were consistent. Model theory then provided a means to prove consistency of axiomatizations relative to Kripke models.

The other direction of the relation between axiomatic descriptions of a logic and semantical ones, is about 'completeness' of the logic.

Weak completeness: if $\models \varphi$ then $\vdash \varphi$

In another, but equivalent formulation, weak completeness says that 'every consistent formula is satisfiable'. To see this, in the above formulation of weak completeness, first take the contraposition, then substitute $\neg \varphi$ for φ , use the logical equivalence $\neg \neg \varphi \leftrightarrow \varphi$ and use the semantics of negation. Weak completeness only says that the validities of some semantic description of a modal logic are also theorems of a corresponding axiomatic description. But, what about logical consequence? That, is, do we also have that if $\Phi \models \varphi$ then $\Phi \vdash \varphi$? To see the answer first recall that for our *local* version of modal logic consequence, we have the deduction theorem saying: $\Phi \models \varphi$ if and only if $\models (\bigwedge \Phi) \rightarrow \varphi$. It follows that for weak consequence and for *finite* sets Φ we have that weak completeness implies completeness of logic consequence. For the stronger notions of consequence we need stronger deductive systems. But if Φ is infinite, we cannot use the deduction theorem. To prove that completeness holds for logic consequence over possibly infinite sets, we have to prove *strong* completeness.

Strong completeness: if $\Phi \models \varphi$ then $\Phi \vdash \varphi$ for arbitrary, and thus possibly *infinite* sets Φ .

An equivalent formulation of strong completeness is that for every set of formulas consistency implies satisfiability, where satisfiability of a set of formulas is defined as existence of a model satisfying all formulas in the set.

A common approach to proving (strong) completeness is by construction of canonical models. A canonical model for a logic is one particular model in where every consistent formula can be satisfied. A closely related approach to proving completeness is to define just a canonical *construction* for building a satisfying model from any given arbitrary consistent formula. But here we only sketch the approach that builds a canonical model for the whole logic.

First of all the worlds of a canonical model are build from maximal consistent sets (MCSs) of formulas of the logic as described by the language and the axiomatization. At first sight this may seem a strange thing to do; until now we stressed that we can use possible worlds to *interpret* modal formulas, and now we construct possible worlds *using* modal formulas. So, syntactic objects become, or construct, semantic ones. But, maybe the reader recalls 'tricks' like these from first order logic (Herbrand models) or algebra (term models).

A set of formulas Φ is maximally L -consistent (is an MCS in the logic L) if: Φ is L -consistent, and no set $\Phi \cup \{\varphi\}$, with $\varphi \notin \Phi$, is L -consistent.

It is a property of canonical models that all elements of L (i.e. all theorems) belong to any maximally consistent set. In particular, all theorems of K belong to all maximally K -consistent sets. Lindenbaum's Lemma says that every L -consistent set Φ can be extended to a maximally L -consistent set Φ_∞ .

The second step is to build an appropriate relational structure and an evaluation of atomic propositions for the canonical model such that any consistent set of formulas is satisfied in some world of the canonical model. Both atomic valuations V^C and the relation R^C are defined in terms of formula membership over worlds w_Φ^C :

$M^C = \langle W^C, R^C, V^C \rangle$ is defined by:

$$W^C = \{w_\Phi^C \mid \Phi \text{ is an MCS}\}$$

$$V^C(p) = w_\Phi^C \text{ if and only if } p \in \Phi$$

$$w_\Phi^C R^C w_\Psi^C \text{ if and only if } \forall \varphi : \Box \varphi \in \Phi \text{ implies } \varphi \in \Psi.$$

The central lemma for proving completeness is the so-called 'Truth Lemma'.

The Truth Lemma: $M^C, w_\Phi^C \models \varphi$ if and only if $\varphi \in \Phi$.

Satisfying the truth lemma ensures that every (possibly infinite!) consistent set is satisfiable. The proof of the truth lemma has an easy and a difficult direction. In this reading companion, we will do neither.

For logics stronger than K , we can extend the canonical models approach. E.g., to prove completeness of $K \cup \{\Box_i \varphi \rightarrow \Box_i \Box_i \varphi\}$ with respect to the class of transitive frames, we only have to prove that the canonical model relative to this system is transitive (since then, each consistent set is satisfied in a model of the right kind).

Based on this observation on the extendibility of the canonical model approach, Sahlqvist formulated a theorem about the connection between correspondence theory and completeness. In particular, Sahlqvist determined a general class of formulas C , for which it holds that any schema γ whose instantiations are in C (we say 'it is in Sahlqvist form'), corresponds to a first-order definable class of frames, whose logic can be axiomatized by adding γ to the axiomatization of the basic language. The value of this theorem is best explained by an example. Since we know that the axiom schema $\Box_i \varphi \rightarrow \Box_i \Box_i \varphi$ corresponds to the first-order condition $\forall s, t, u \in W : sRt \wedge tRu \rightarrow sRu$ on frames, the theorem says that the logic of the class of transitive frames is axiomatized by $K \cup \{\Box_i \varphi \rightarrow \Box_i \Box_i \varphi\}$. Sahlqvist forms are defined as follows.

A boxed atom is a propositional atom preceded by a number (possibly 0) of boxes, i.e. a formula of the form $\Box \dots \Box p$.

A Sahlqvist antecedent is a formula constructed using \wedge , \vee , and \Diamond_i from boxed atoms, and negative formulas (including the constants \perp , \top).

A Sahlqvist implication is a formula $A \rightarrow B$, where A is a Sahlqvist antecedent, and B is a positive formula.

A Sahlqvist formula is constructed from Sahlqvist antecedents using \wedge and \Box_i (unlimited), and using \vee on formulas with no common variables.

The Sahlqvist class is an extremely useful concept for proving completeness of modal logics. But of course, there are many cases where we cannot use it. Also, the Sahlqvist forms are still being extended to incorporate new formulas. It thus appears that the relation between correspondence theory and completeness is still not fully understood.

1.5 Example of combining modal logics: products

There are many ways in which separate modal logics can be combined in one logical system. Here we only discuss *products*. Product logics are characterized by product frames. A product frame is a frame where the set of worlds have a multi-dimensional structure. The formal definition is as follows.

Given two frames (U_0, R_0) and (U_1, R_1) , their *product frame* $(U_0 \times U_1, R_H, R_V)$, with $U_0 \times U_1$ the Cartesian product of the worlds from the originating models, and R_H and R_V accessibility relations defined by (1) $(w, x)R_H(y, z)$ if and only if wR_0y and $x = z$, and (2) $(w, x)R_V(y, z)$ if and only if xR_0z and $w = y$.

The first order conditions characterizing product frames are as follows.

Commutativity: for all d, e, f such that $dR_H eR_V f$ there is a g such that $dR_V gR_H f$, and for all d, e, f such that $dR_V eR_H f$ there is a g such that $dR_H gR_V f$

Confluence: for all d, e, f such that $dR_H e$ and $dR_H f$, there is a g such that $eR_H g$ and $fR_H g$

The corresponding modal axiom schemas are:

Commutativity: $HV\varphi \leftrightarrow VH\varphi$

Confluence: $\neg H\neg V\varphi \rightarrow V\neg H\neg\varphi$

Given two normal model logics L_1 and L_2 over the frame classes \mathcal{F}_1 and \mathcal{F}_2 we denote the logic over their product frames by $L_1 \times L_2$. Product logics arise naturally in many applications of multi-modal logic to the multi-agent domain. We will encounter them in section 4.4.1. But they are known for their high computational complexities (see section 1.6).

1.6 Decidability and complexity issues

While Sahlqvist's Theorem provides a general completeness result, no such general results exist in what concerns decidability of satisfiability of modal formulas in a given modal logic, as well as the complexity of the decision problem. For example, for most of the logics defined by Sahlqvist forms decidability is unknown. We here recall the main results for the basic modal logics.

1.6.1 Decidability

Given a modal logic L , the satisfiability problem for L is decidable if there exists an algorithm which for every input formula φ decides whether φ is satisfiable in L or not. Decidability of the validity problem for L and the consequence problem for L are defined analogously.

Decidability is usually proved by the filtration method, where the canonical model for the logic L is reduced by restricting attention to the set of subformulas Γ_φ of the formula φ under concern: if two possible worlds w_1 and w_2 of M^C satisfy exactly the same elements of Γ_φ then they are identified. In this way a finite model is obtained, in which satisfiability of φ can be decided.

In this way it can be shown that all the normal modal logics defined by any combinations of the axiom schemas D, T, B, 4, 5 are decidable.

Decidability proofs for more 'delicate' logics such as LTL, PDL and epistemic logic with common knowledge are often done by subtle modifications of the above method, such as the so-called Fischer-Ladner closure.

While modal logics are 'surprisingly often decidable', there also exist numerous modal logics that are undecidable. An example is the product logic $S5 \times S5 \times S5$.

1.6.2 Complexity

Suppose L is a decidable logic. Roughly speaking, the complexity of L is the difficulty of the decision algorithm. Difficulty is measured by the amount of time and/or space required by the algorithm, as a function of the input size.

The decidable modal logics cover a wide range of complexity classes, ranging from NP up to EXPSPACE. We recall the hierarchy that is relevant for us:

$$\text{NP} \subset \text{PSPACE} \subset \text{EXPTIME} \subset \text{NEXPTIME}$$

Satisfiability in S5 can be decided in nondeterministic polynomial time, i.e. the satisfiability problem is in NP. This upper bound is clearly tight, given that satisfiability of classical propositional logic is already NP-hard: the satisfiability problem for S5 is NP-complete. The same holds for its weakenings KD45 and K45.

Satisfiability in the basic modal logics between K and S4 can be decided in polynomial space (PSPACE). This upper bound is tight, i.e. satisfiability is PSPACE complete. By basic modal logics between K and S4 we mean the logics K, KD, KT, K4, KD4, and KT4, alias S4. Satisfiability in $S5_n$ (the multimodal version of S5) is also PSPACE-complete. Satisfiability in linear temporal logic LTL is also PSPACE-complete (see Section 1.8.5).

Logical consequence in K can be decided in (deterministic) exponential time: it is EXPTIME-complete. This also holds for PDL and for epistemic logic with common knowledge $S5_n^C$ (see Section 1.8.2).

Satisfiability in the product logic $S5 \times S5$ is decidable in nondeterministic exponential time: it is NEXPTIME-complete. Note that an NEXPTIME upper bound can generally be obtained from a decidability proof that is done via filtration.

1.7 Non-normal modal logics: neighborhood models

All modal logic discussed so far were *normal*. This means that they all validate the K-axiom, and that they all can be given semantics using Kripke structures. In this section we briefly explain how we can give semantics to logics that are weaker than K. To be very clear: we do not change the language of modal logic. The only thing we want to drop is the axiom $\Box_i(\varphi \rightarrow \psi) \rightarrow (\Box_i\varphi \rightarrow \Box_i\psi)$. We will call such logics *weak* modal logics.

Since by throwing out $\Box_i(\varphi \rightarrow \psi) \rightarrow (\Box_i\varphi \rightarrow \Box_i\psi)$, we have thrown out the *bottom* of the type of modal logics we discussed so far, we need a new bottom. This bottom consists in assuming that for weak modal logics, we have at least that the modal box preserves logical equivalence. That is, weak modal logics satisfy, apart from the propositional axioms and rules, the following rule.

(Logical equivalence) from $\varphi \leftrightarrow \psi$ infer $\Box_i\varphi \leftrightarrow \Box_i\psi$

Of course, for normal modal logics this rule is also valid: it is an instance of the well-known rule of (not necessarily uniform) replacement of logically equivalent formulas.

However, how do we give semantics to such weak languages? We can no longer use Kripke models, since these correspond to the K-axiom we no longer have.

One possibility is to consider a specific kind of generalization of Kripke models called ‘neighborhood models’. In a neighborhood model we still have possible worlds, but the binary relation is now generalized to a mapping from worlds to sets of *sets* of worlds. The sets of worlds reachable from a certain world are called ‘neighborhoods’. One can now view the traditional Kripke models as particular neighborhood models where the neighborhoods are singleton sets. Formally, a neighborhood model is defined as follows.

A neighborhood model is a triple $M = \langle W, N, V \rangle$, where

W is a non-empty set of possible worlds,

N is a function $N : Labels \times W \longrightarrow 2^{2^W}$ mapping each modal label and world to a set of *sets* of possible worlds,

V is a function $V : W \longrightarrow 2^{ATM}$ mapping each world to a subset of atomic proposition symbols.

These new models require a new truth definition for ‘modal’ formulas.

Given a neighborhood model $M = \langle W, N, V \rangle$, a formula of the form $\Box_i \varphi$ is said to be true with respect to a world w in W , notation $M, w \models \Box_i \varphi$, if $\exists N \in N(i, w)$ such that $N = \{w' \mid M, w' \models \varphi\}$.

The rule of closure under substitution of logical equivalents is sound and complete with respect to the semantics resulting from this truth definition. This is not too surprising, given that the neighborhood semantic interpretation of $\Box_i \varphi$ says nothing more than that the group of φ worlds is reachable (as a set).

Starting from the lower bound to the logic given by the above axiomatization and semantics we can slowly strengthen the logic. Although there is no general correspondence theory for neighborhood models, it is for instance easy to see that closure of reachability of neighborhoods under intersections results in $\Box_i \varphi \wedge \Box_i \psi \rightarrow \Box_i(\varphi \wedge \psi)$. With both closure under reachability of intersections and closure under reachability of supersets (i.e., super-neighborhoods) we are back at normal modal logic.

1.8 Modal logic: application to some reasoning domains

The mathematical theory of modal logic can be applied to many reasoning domains. Any domain suited for abstract representation as a binary relation over possible worlds, qualifies.

1.8.1 Dynamic Logic

In dynamic logic [Pra76, HKT00], the worlds are abstract representations of states of some computer system, and the binary relations represent system transitions. The logic’s central modal operators are $[\alpha]\varphi$ and $\langle \alpha \rangle \varphi$ for ‘after any execution of programm α , φ holds’, and ‘an execution of programm α resulting in a system state where φ , is possible’, respectively. The procedures / programs α are confined to regular languages over a set of atomic action letters plus a test procedure, and sometimes the converse operation:

$$\alpha, \beta, \dots ::= a \mid \alpha \cup \beta \mid \alpha; \beta \mid \alpha^* \mid \varphi? \mid \alpha^-$$

Atomic actions are interpreted in the standard modal way by multi-modal Kripke structures where indexes of accessibility relations correspond to atomic action names of the regular action language. Special to Dynamic logic is the interpretation of the compound actions. They get a relational interpretation like in relation algebra [Tar41]. That is, the $;$ is interpreted as concatenation of relations, $*$ as reflexive transitive closure, etc.

Although originally restricted to the above described modal logic over regular event languages, the label ‘dynamic logic’ is now generally used for any modal logic over explicit action languages.

1.8.2 Epistemic Logic

In the application of the modal theory to so called epistemic logic, we can take the notion of a possible world directly to mean that a world is considered possible according to an agents knowledge or belief. For example, if an agent knows that the earth is round, in all worlds he considers possible, the earth is round. And if an agent does not know whether or not it rains outside, his knowledge allows for at least two contingencies: one where it rains and one where it does not rain. A relation between two worlds thus represents whether from the viewpoint of one world, the other world is an epistemic possibility.

The modal operator $Bel_i\varphi$, with i ranging over a set of agents AGT then reads as ‘agent i believes that φ ’, and the ‘diamond’ operator $\neg Bel_i\neg\varphi$ as ‘ i ’s beliefs allow for the possibility that φ ’.

The main difference between ‘knowledge’ and ‘belief’ is that for knowledge the ‘truth axiom’ is usually assumed $Bel_i\varphi \rightarrow \varphi$. The truth axiom is nothing more than the modal axiom of reflexivity. For ‘belief’ it is not assumed, to allow for the possibility that what is actually the case (semantically represented by the present world) is different from what the agent believes (all worlds accessible from the present world, that is, the worlds possible according to the agent’s belief).

Well-known properties for knowledge and belief that can suitably be expressed as modal axiom schemas are:

Introspection (transitivity): $Bel_i\varphi \rightarrow Bel_iBel_i\varphi$.

Negative introspection (Euclidicity): $\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi$.

1.8.3 STIT logic

The starting point for any STIT semantics is that *acting* is the same as *ensuring* the actual world is among a set of possible worlds that satisfy the property being secured by the action. For instance, ‘ A closes the door’ is interpreted as ‘ A ensures that in all possible worlds resulting from his action, the door is closed’. If i ranges over a set AGT of agents, roughly, the STIT expression $[i]\varphi$ for ‘agent i sees to it that φ ’ means that due to the action performed by i , the actual world is among the ones satisfying φ . It follows immediately, that one of the central axioms of STIT theory is the so called ‘success axiom’: $[i]\varphi \rightarrow \varphi$. STIT semantics captures many more such principles of agency, and in section 4 we will give them in the form of axioms and their corresponding semantic conditions.

1.8.4 Deontic Logic

Deontic logic is about notions such as ‘permission’, ‘prohibition’ and ‘obligation’. Obligations and prohibitions can be violated, and in deontic logic one aims at modeling reasoning about violations, or their absence, without taking violations as inconsistencies. Any reasonable deontic logic should thus be able to represent violations.

Von Wright, in 1951, single-handedly started this area of logic by publishing a paper [Wri51] where he observed that the notion of ‘obligation’ behaves like a modal box, and ‘permission’ like a modal diamond. Basically, Von Wright suggested the modal logic KD for the notion of obligation. Permission is then the modal dual of obligation, and prohibition the negation of permission. Violations against an obligation can be consistently represented in this logic as $O_i\varphi \wedge \neg\varphi$, with i ranging over a finite set of agents AGT . The D-axiom says that obligations do not conflict (often one speaks of ‘consistency’ of the obligations), that is, formulas of the form $O_i\varphi \wedge O_i\neg\varphi$ are not satisfiable.

Chisholm in his famous note [Chi63] where he describes the paradox that inherited his name, argued that Von Wright’s standard deontic logic is not suited for contrary to duty reasoning. To model this type of reasoning, one needs a notion of *conditional* obligation that cannot be captured directly by a standard normal modal logic.

1.8.5 Linear Time Temporal Logic

Modal logic can also be suitably used to model reasoning about time. In this application, possible worlds represent moments, and the accessibility relation models the flow of time. The language of linear time temporal logic has the following syntax.

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \psi \mid G\varphi \mid X\varphi \mid \varphi\mathcal{U}\psi$$

The operator G stands for ‘henceforth’, X for ‘next’, and \mathcal{U} for ‘until’.

The Kripke models of LTL are infinite time paths into the future: W can be supposed to be the set of positive integers $\{0, 1, 2, \dots\}$. The next operator refers to next elements of the path, the henceforth

operator to all future moments on the path, and the until operator to some future moment where ψ , and the moments until then, where φ . The formal semantics of ‘until’ is thus as follows.

$M, n \models \varphi\mathcal{U}\psi$ if there is a world $n_\psi \geq n$ such that $M, n_\psi \models \psi$, and $M, m \models \varphi$ for every $m \in W$ such that $n \leq m < n_\psi$.

In our exposition of the notion of ‘intention’, we also need the Before operator. $\varphi\mathcal{B}\psi$ is defined to be an abbreviation of $\neg\psi\mathcal{U}\varphi$.

In LTL one can express so called safety, liveness, and fairness properties. This classification [AS87] is from the area of verification of concurrent systems. Roughly, safety properties, saying that it is preserved over time that something bad will not happen, have the form $G\neg\varphi$, while liveness properties, saying that starting at any future moment some good thing will happen eventually, have the form $G\neg G\neg\varphi$.

LTL has been given a complete axiomatization. But since this axiomatization is not as straightforward as one might expect from a modal logic, we do not give it here.

2 Logical Frameworks for Multi-Agent Systems: Cohen & Levesque

2.1 Motivation and background

The 1990 formalization of Bratman’s theory of intention [Bra87] by Cohen and Levesque (C&L henceforth) [CL90] was designated one of the most influential papers in the domain at AAMAS’06. Their approach is based on a logical framework integrating the concepts of belief, event and action, time, and preference. In that framework they successively define several notions of goal, and finally intention.

Intention is thus a non-primitive mental attitude that is defined by means of modal operators of belief, preference, and time.

In this section we sketch the framework of the 1990 paper by Cohen and Levesque. We enhance their account by integrating more recent progress in logics of belief and action, following [HL04]. Among the others to give a formal notion of intention based on Bratman were Meyer [MHL99], Rao and Georgeff [RG95] and Wooldridge [Woo02].

2.2 Syntax

There is a set of *events* $EVT = \{\alpha_1, \alpha_2, \dots\}$ and a set of *agents* $AGT = \{i, j, \dots\}$. Actions are events that are brought about by agents.

We have a standard multimodal language made up of the following:

- a standard modal operator of belief Bel_i , where $i \in AGT$ is an agent (see Section 1.8.2)
- a temporal modal operator G (see 1.8.5)
- dynamic logic operators $[a]$, where $a \in EVT$ is an event (see 1.8.1)
- a modal operator of preference $Pref_i$, where $i \in AGT$ is an agent

We recall that $G\varphi$ reads “ φ holds henceforth”, its dual $F\varphi$ reads “ φ will eventually hold”, $X\varphi$ reads “ φ will hold at the next time point”. The new operator $Pref_i\varphi$ reads “ i chooses that φ ”, or “ i prefers that φ ”, or “ i wants φ to be true”.

An agent’s intention-that φ , written $Int_i\varphi$, is a non-primitive concept, to be defined by means of the other modal operators.

2.3 Semantics

2.3.1 The ingredients

Time is supposed to be deterministic. Hence reading of the PDL formula $[a]\varphi$ simplifies to “if α happens then φ holds after α ”. The dual $\langle a \rangle\varphi = \neg[a]\neg\varphi$ therefore expresses that α happens and φ is true afterwards. Hence $[a]\perp$ expresses that α does not happen, and $\langle a \rangle\top$ expresses that α happens.

To speak about sequences of more than one event there is an LTL operator G : $G\varphi$ expresses that henceforth φ holds (see Section 1.8.5). The dual operator F is defined by $F\varphi = \neg G\neg\varphi$ (‘eventually φ ’).

The logic of the belief operator Bel is KD45 (see Section 1.8.2).

Among all the worlds in $\mathcal{B}_\gamma(\sqsupseteq)$ that are possible for i , there are some that i prefers. C&L say that i chooses some subset of $\mathcal{B}_\gamma(\sqsupseteq)$. Semantically, these worlds are accessible via yet another relation \mathcal{C}_γ . $Pref_i\varphi$ expresses that agent i chooses that φ . We sometimes also say that i prefers that φ .¹ Without surprises, we have the standard truth condition for modal operators: $w \models Pref_i\varphi$ if $w \models \varphi$ for every $w' \in \mathcal{C}_\gamma(\sqsupseteq)$.

- \mathcal{C}_γ is serial, transitive, and euclidian.

¹While C&L use a modal operator ‘goal’ (probably in order to have a uniform denomination w.r.t. the different versions of goals they study), it seems more appropriate to us to use the term ‘choice’ or ‘preference’.

It will follow from the sequel that the set of an agent's choices is nonempty.

What is the relation between choice and belief? As said above, an agent only chooses worlds he considers possible:

- $\mathcal{C}_\gamma(\sqsupset) \subseteq \mathcal{B}_\gamma(\sqsupset)$.

Choice is thus a mental attitude that is weaker than belief: belief implies choice. It is moreover required that worlds chosen by i are also chosen from i 's possible worlds, and vice versa:

- if $w \in \mathcal{B}_\gamma(\sqsupset')$ then $\mathcal{C}_\gamma(\sqsupset) = \mathcal{C}_\gamma(\sqsupset')$.

Such a semantics validates the equivalences

$$\begin{aligned} \text{Pref}_i \varphi &\equiv \text{Bel}_i \text{Pref}_i \varphi \\ \neg \text{Pref}_i \varphi &\equiv \text{Bel}_i \neg \text{Pref}_i \varphi \\ \text{Pref}_i \varphi &\equiv \text{Pref}_i \text{Pref}_i \varphi \\ \neg \text{Pref}_i \varphi &\equiv \text{Pref}_i \neg \text{Pref}_i \varphi \end{aligned}$$

Suppose the actual world is w , and some event α occurs leading to a new actual world w' . Which worlds are possible for agent i at w' ? According to Moore [Moo85] and Scherl and Levesque [SL93, SL03], i makes 'mentally happen' α in all his worlds $v \in \mathcal{B}_\gamma(\sqsupset)$, and then collects the resulting worlds $\mathcal{R}_\alpha(\sqsupset)$ to form the new belief state $\mathcal{B}_\gamma(\sqsupset') = \mathcal{R}_\alpha(\mathcal{B}_\gamma(\sqsupset)) = \bigcup_{\sqsupset \in \mathcal{B}_\gamma(\sqsupset)} \mathcal{R}_\alpha(\sqsupset)$. This identity must be restricted in order to keep i 's beliefs consistent, i.e. to avoid $\mathcal{B}_\gamma(\sqsupset') = \emptyset$. We thus obtain:

- If $w \in \mathcal{R}_\alpha(\sqsupset')$ and $\mathcal{R}_\alpha(\mathcal{B}_\gamma(\sqsupset)) \neq \emptyset$ then $\mathcal{B}_\gamma(\sqsupset') = \mathcal{R}_\alpha(\mathcal{B}_\gamma(\sqsupset))$.

This relies on the hypothesis that events are uninformative: apart from the mere occurrence of α he should learn nothing about α 's particular effects that hold in w' , and $\mathcal{B}_\gamma(\sqsupset')$ only depends on $\mathcal{B}_\gamma(\sqsupset)$ and α .

How do an agent's choices evolve? We recall that to each possible world there is associated a sequence of events (its history). Therefore agent i 's choices concern not only possible states of the world, but also possible histories. We therefore suppose that i 's preferences after α are just the images by α of its preferred worlds before α . Just as for belief, this identity must be restricted in order to keep i 's choices consistent. We thus obtain the constraint:

- If $w \in \mathcal{R}_\alpha(\sqsupset')$ and $\mathcal{R}_\alpha(\mathcal{C}_\gamma(\sqsupset)) \neq \emptyset$ then $\mathcal{C}_\gamma(\sqsupset') = \mathcal{R}_\alpha(\mathcal{C}_\gamma(\sqsupset))$.

Again, note that such an explanation is in accordance with our hypotheses.

Our conditions say nothing about i 's beliefs after a surprising action occurrence, i.e. when $\mathcal{R}_\alpha(\mathcal{B}_\gamma(\sqsupset)) = \emptyset$. In this case i must revise his beliefs. Integrations of belief revision into a logic of action and belief have been proposed in [SPLL00] and [HL02].

Our conditions do not constrain either i 's choices when $\mathcal{R}_\alpha(\mathcal{B}_\gamma(\sqsupset)) = \emptyset$, i.e. after an unwanted action occurrence. Then i has to revise his choices. There are two cases. First, if $\text{Pref}_i[a] \perp$ and $\text{Bel}_i[a] \perp$ then a surprising event has occurred, and the agent has to revise both his beliefs and his choices. In the second case we have $\text{Pref}_i[a] \perp$ and $\neg \text{Bel}_i[a] \perp$. Then i did not believe the event was impossible, but preferred so. Devices such as a preference relation have to be integrated here [Tho00].

2.3.2 The models

We have defined the semantics of a basic logic of action, belief, and choice. Models have the form $\langle W, \mathcal{B}, \mathcal{C}, \mathcal{R}, \mathcal{R}_G, \mathcal{V} \rangle$, where W is a set of possible worlds, \mathcal{B} and \mathcal{C} associate accessibility relations to every agent, \mathcal{R} associates an accessibility relation to every action, \mathcal{R}_G is the accessibility relation for G , and \mathcal{V} associates a valuation to every possible world. We call *C&L models* the set of models satisfying all the constraints imposed in the preceding sections and write $\models_{\text{C\&L}} \varphi$ if φ is valid in C&L models. We write $\mathcal{S} \models_{\text{C\&L}} \varphi$ if φ is a logical consequence of the set of formulas \mathcal{S} in C&L models.

2.4 Achievement goal, persistent goal and intention

An agent i has the **goal** that φ if φ holds in the future of all of i 's preferred histories: $Pref_i F\varphi$. Then an **achievement goal** of agent i is a goal of i of which i believes it is not achieved yet:

$$AGoal_i^{CL}\varphi \stackrel{\text{def}}{=} Pref_i F\varphi \wedge Bel_i \neg\varphi$$

A **persistent goal** is an achievement goal that persists until it is either achieved, or believed to be impossible:

$$PGoal_i^{CL}\varphi \stackrel{\text{def}}{=} AGoal_i^{CL}\varphi \wedge Before_{(Bel_i\varphi \vee Bel_i G\varphi)} \neg Pref_i F\varphi$$

where $Before_{(Bel_i\varphi \vee Bel_i G\varphi)} \neg Pref_i F\varphi$ is the LTL formula expressing that $Bel_i\varphi \vee Bel_i G\varphi$ is true before $\neg Pref_i F\varphi$. It has been proved in [HL04] that persistence of achievement goals can be *deduced* from our no forgetting principle for choice as long as the event is not unwanted:

Theorem 1 ([HL04]). $\models_{C\&L} (AGoal_i\varphi \wedge \neg Pref_i[a]\perp) \rightarrow [a](AGoal_i\varphi \vee Bel_i\varphi)$

We inherit the properties of achievement goals concerning logical principles, the side effect problem, and persistence. C&L's original definition is that a persistent goal that φ is an achievement goal that φ that can only be abandoned if (1) φ is achieved, or (2) the agent learns that φ can never be achieved, or (3) for some other reason. This leads to the principle $PGoal_i\varphi \rightarrow [a](PGoal_i\varphi \vee Bel_i\varphi \vee Bel_i G\neg\varphi \vee \psi)$, where ψ is an unspecified condition accounting for case (3). Theorem 1 makes (3) more precise by identifying it with the occurrence of an unwanted event, which is the only case when achievement goals have to be revised.² Indeed, the theorem tells us that C&L's case (2) is excluded when $\neg Pref_i[a]\perp$ holds: in this case we are guaranteed that i will not learn through α that φ will be false henceforth. Given our hypothesis that events are uninformative, this is as it should be.

Finally, C&L define an **intention that** φ as a persistent goal to the achievement of which the agent actively contributes, in the sense that in every preferred history there must be some action α whose author is i and which brings about φ . Noting $i:\alpha$ such an action and using quantification over actions this can be written:

$$Int_i^{CL}\varphi \stackrel{\text{def}}{=} PGoal_i\varphi \wedge Pref_i F\exists i:\alpha \langle i:\alpha \rangle \varphi$$

where $\langle i:\alpha \rangle \varphi$ reads “ i does action α and φ holds after α 's occurrence”.

2.5 Discussion of C&L

First of all, C&L's definition of intention puts some technical problems: how can we define quantification over actions in PDL?

But it can also be criticized on philosophical grounds. First, it is too strong in some respect [Sad00]: it is not necessarily the *last* action of mine whose effect is φ , because there might be another agent making φ true because I have asked him to do so at a previous stage.

Second, C&L's definition is *too weak* because it lacks a causal connection between the agent and the goal. Indeed, suppose that agent i wants to go to a shoe shop on Saturday:

$$Pref_i F \langle i:GoToShoeShop \rangle \top$$

Moreover suppose that i has the persistent goal that for all weekend it will be sunny: $PGoal_i^{CL} Sunny$. The previous preference and the persistent goal are completely unrelated, that is, *going to a shoe shop on Saturday* is not part of a plan for achieving the result *for all weekend it will be sunny*. Agent i

²In the case where i is the agent of α (noted $i:\alpha$) one might reasonably suppose that $Pref_i[i:\alpha]\perp \rightarrow [i:\alpha]\perp$, i.e. there are no such unwanted action occurrences. We then get unconditioned persistence of achievement goals: $AGoal_i\varphi \rightarrow [i:\alpha](AGoal_i\varphi \vee Bel_i\varphi)$. This is related to intentional actions as discussed in C&L's [CL90, section 4.2.1], where moreover $Bel_i[i:\alpha]\perp \vee Bel_i\neg[i:\alpha]\perp$ is assumed; see [LHC06]. We just note that such principles are of the Sahlqvist type, and can be added to the logic without harm.

is simply endorsing two different goals at the same time. From the fact that i wants to perform the action of going to a shoe shop on Saturday and the fact that he has the persistent goal that for all weekend it will be sunny it follows that each of i 's preferred histories has the action *going to a shoe shop* of i leading to a state where the fact *Sunny* holds: $Pref_i F \langle i:GoToShoeShop \rangle Sunny$. From the previous it follows that each of i 's preferred histories has some action of i leading to a state where the fact *sunny* holds: $Pref_i F \exists i:\alpha \langle i:\alpha \rangle Sunny$. According to Cohen and Levesque's definition of *intention to be*, i has the intention that *it will be sunny*: $Int_i^{CL} Sunny$. This consequence seems unacceptable. There is common agreement in philosophy that we cannot reasonably say that we intend that some event occurs when we believe that the occurrence of this event is independent of us. According to Searle for instance I cannot say that I intend that it will rain or I intend that the sun will rise etc... [Sea83].

3 The power of cooperation: Coalition Logic

3.1 Motivation

Coalition Logic (CL) was proposed by Pauly in 2001. It is perhaps the simplest logic to reason about abilities of agents and groups of agents. It is about sentences of the form “agent i has the ability to achieve φ ” and “group J has the ability to achieve φ ”, where ability is understood as an appropriate choice in i ’s repertoire in the individual case, and as an appropriate choice of J ’s members in the group case, that is supposed to be simultaneous. The latter is thus about a particular form of joint action. J is also called a coalition, although CL does not provide linguistic resources to speak about other ingredients of coalitions (such as J ’s internal organization or its goals).

The central hypothesis of CL is that agents are completely independent in their choice: whatever action they go for, its execution will not be blocked by the other agents, and the respective choices of all the agents can be combined to obtain the outcome. This is called the *superadditivity constraint*.

There are two different semantics for CL: in game models it is supposed that a group J ’s abilities are completely determined by that of its members, while (roughly speaking) in neighborhood models a coalition may be stronger than the sum of its members.

3.2 Syntax

We have a standard multimodal language with modal operators $\langle\!\langle J \rangle\!\rangle \mathbf{X}$, where J is a set of agents. $\langle\!\langle J \rangle\!\rangle \mathbf{X}\varphi$ is read “coalition J is able to ensure φ ”. Our syntax seems quite more elaborate than Pauly’s own syntax, which is simply $[J]\varphi$ for the same operator. But we want to use that syntax for STIT. Furthermore, our syntax $\langle\!\langle J \rangle\!\rangle \mathbf{X}$ hints to the fact that the semantics of the central coalition logic operator hides two modal quantifications (the square and the sharp brackets) and a temporal ‘next’ operation. In the sequel we will indeed decompose the CL operator into these modal constituents. But for the moment we have to view the operator $\langle\!\langle J \rangle\!\rangle \mathbf{X}$ as a ‘monolith’.

3.3 Semantics: neighborhood models

Neighborhood models for coalition logic were proposed in [Pau02]. They are neighborhood models (cf. Section 1.7) satisfying moreover a superadditivity constraint.

Definition 1 (effectivity function). *Given a set of agents AGT and a set of states S , an effectivity function is a function $E : 2^{AGT} \longrightarrow 2^{2^S}$. An effectivity function is said:*

- *J -maximal iff for all $X \subseteq S$, if $S \setminus X \notin E(\bar{J})$ then $X \in E(J)$.*
- *outcome monotonic iff for all $X \subseteq X' \subseteq S$ and for all $J \subseteq AGT$, if $X \in E(J)$ then $X' \in E(J)$.*
- *superadditive iff for all X_1, X_2, J_1, J_2 such that $J_1 \cap J_2 = \emptyset$, $X_1 \in E(J_1)$ and $X_2 \in E(J_2)$ imply that $X_1 \cap X_2 \in E(J_1 \cup J_2)$.*

E intuitively associates every coalition J to a set of $X \subseteq S$ (a set of possible outcomes) for which J is effective. That is, J can force the world to be in some state of X at the next step.

Definition 2 (playable effectivity function). *An effectivity function $E : 2^{AGT} \longrightarrow 2^{2^S}$ is said playable iff*

1. $\forall J \subseteq AGT, \emptyset \notin E(J)$; (Liveness)
2. $\forall J \subseteq AGT, S \in E(J)$; (Termination)
3. E is AGT -maximal;
4. E is outcome-monotonic; and

5. E is superadditive.

Definition 3. A coalition model is a pair $((S, E), V)$ where:

- S is a nonempty set of states;
- $E : S \longrightarrow (2^{AGT} \longrightarrow 2^{2^S})$ is a playable effectivity structure;
- $V : S \longrightarrow 2^{Prop}$ is a valuation function.

The mapping E associates every state s to a playable effectivity function $E(s)$. We will write $E_s(J)$ instead of $E(s)(J)$.

Truth conditions are standard for classical formulas. We evaluate the coalitional operators against a coalition model M and a state s as follows:

$$M, s \models \langle\!\langle J \rangle\!\rangle \mathbf{X}\varphi \text{ iff } \{s \mid M, s \models \varphi\} \in E_s(J).$$

3.4 Semantics: game structures

In [Pau02], Marc Pauly investigates the link between coalition models and strategic games.

Definition 4. A strategic game is a tuple $G = (S, \{\Sigma_i \mid i \in AGT\}, o)$ where S is a nonempty set, Σ_i is a nonempty set of choices for every agent $i \in AGT$, $o : \prod_{i \in AGT} \Sigma_i \longrightarrow S$ is an outcome function which associates an outcome state in S with every combination of choice of agents (choice profile).

It appears that there is a strong link between a coalition model (whose effectivity structure is *playable* by definition) and a strategic game.

Definition 5. Given a strategic game G , the effectivity function $E_G : 2^N \longrightarrow 2^{2^S}$ of G is defined as $X \in E_G(C)$ iff there is $\sigma_C \in \prod_{i \in C} \Sigma_i$ such that for every $\sigma_{\bar{C}} \in \prod_{i \in \bar{C}} \Sigma_i$ we have $o(\sigma_C \times \sigma_{\bar{C}}) \in X$.

Pauly then gives the following characterization:

Theorem 2 ([Pau02]). An effectivity function E is playable iff it is the effectivity function of some strategic game.

Definition 6. Let E be an effectivity function. A set $Y \subseteq S$ is called a minimal effectivity outcome at s for J iff (1) $Y \in E_s(J)$ and (2) there is no $Y' \in E_s(J)$ s.t. $Y' \subset Y$.

Definition 7. The non-monotonic core of E is the mapping $\mu_E : 2^{AGT} \times S \longrightarrow 2^{2^S}$ such that $\mu_E(J, s) = \{Y \mid Y \text{ is a minimal effectivity outcome at } s \text{ for } J\}$.

The outcome of a strategic game is completely determined when every agent has made its choice. The following holds.

Proposition 1. $\mu_E(AGT, s)$ is a set of singletons.

From Definition 5, this is a corollary of Theorem 2.

3.5 Axiomatization

The set of formulas that are valid in coalition models is completely axiomatized by the following principles [Pau02].

- | | |
|-------------|--|
| (ProTau) | all tautologies of classical propositional logic |
| (\perp) | $\neg \langle\!\langle J \rangle\!\rangle \mathbf{X}\perp$ |
| (\top) | $\langle\!\langle J \rangle\!\rangle \mathbf{X}\top$ |

- (N) $\neg\langle\{\emptyset\}\mathbf{X}\neg\varphi \rightarrow \langle\{AGT\}\mathbf{X}\varphi$
- (M) $\langle\{J\}\mathbf{X}(\varphi \wedge \psi) \rightarrow \langle\{J\}\mathbf{X}\psi$
- (S) $\langle\{J_1\}\mathbf{X}\varphi \wedge \langle\{J_2\}\mathbf{X}\psi \rightarrow \langle\{J_1 \cup J_2\}\mathbf{X}(\varphi \wedge \psi)$ if $J_1 \cap J_2 = \emptyset$
- (MP) from φ and $\varphi \rightarrow \psi$ infer ψ
- (RE) from $\varphi \leftrightarrow \psi$ infer $\langle\{J\}\mathbf{X}\varphi \leftrightarrow \langle\{J\}\mathbf{X}\psi$

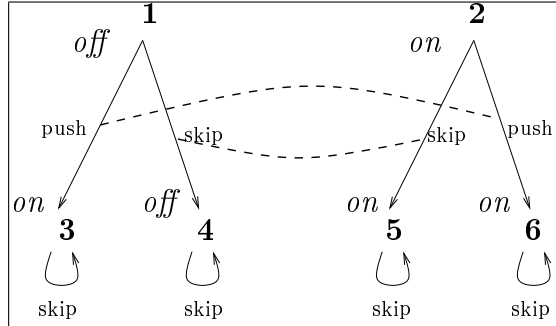
Theorem 3 ([Pau02]). *The principles (ProTau), (\perp), (\top), (N), (M), (S), (MP) and (RE) are complete with respect to the class of all coalition models.*

3.6 Discussion: epistemic extensions and their problems

While CL allows to reason about the abilities and the power of an agent, it does not capture all of its facets; in particular it does not reckon for the epistemic aspect.

The most natural way of adding uncertainty to CL is to augment CL-models by a family of equivalence relations among *states* (one for each agent), interpreting a standard normal S5 operator K_i in the language. Although such an extension of CL looks quite simple and natural, it turns out to be unsatisfactory. To explain this consider two scenarios.

Example 1. *Ann is in a room. She is blind and cannot distinguish a world where the light is off from a world where the light is on. The light in the room is controlled by a button that activates a timer. When the button is pushed the bulb light will shine for a determinate time. When the light is on, there is no way to switch it off. Ann can also do nothing (skip). The actual situation is that the light is off and Ann is pushing the button.*

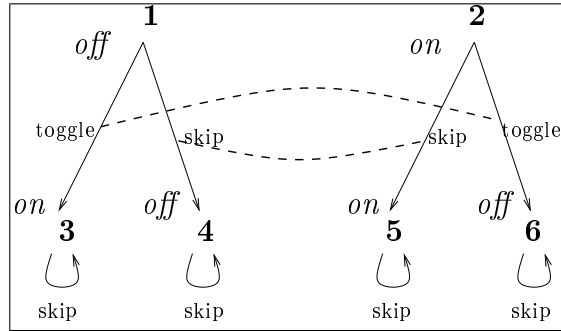


In this example Ann can be said to have the power to achieve that the light is on: Ann has an action (viz. pushing) allowing her to switch the light on, and she knows that.

Example 2. *Ann is in a room. She is blind and cannot distinguish a world where the light is off from a world where the light is on. The light in the room is controlled by a switch. In her repertoire of actions, Ann can toggle or remain passive (skip), which correspond to switching the state of the light and maintaining the state of the light, respectively. The actual situation is that the light is off and Ann toggles.*

In this example, Ann *cannot* truly be said to have the power to switch the light on: she does not know for which action to go, because the choice of the appropriate action depends on the possible world. I.o.w., there is no action in Ann's repertoire of which she knows to achieve *Light*. In the AI planning literature, it is said that there is no *conformant plan* to ensure that the light is on.

Nevertheless, $\text{Know}_{Ann}\langle\{Ann\}\mathbf{X}\text{Light}$ is true in the actual world (whatever it is) of our example model, according to our epistemic extension of CL, suggesting that Ann knows that she has the power to switch the light on.



So, how can CL be extended in order to take into account an agent's uncertainty about the state of the world, and in particular in order to express uniform choices? As we have argued in [BHT07b] it is not only the simple epistemic extension of CL that we have discussed above that leads to problems, but all the known approaches to the problem of uniform strategies in the literature are unlikely to succeed. Note first that in example 2 above we might have given different names to the actions. And there is no reason why this renaming should be uniform. In particular, the *left* toggle action can be called '*put the light on*' and the *right* toggle action '*put the light off*'. Obviously, non-uniform renaming of actions should not influence Ann's basic capabilities or her knowledge concerning her capabilities. However, all CL- and ATEL-based approaches in the literature do not satisfy this consideration. In these variants and extension of ATEL (see e.g. [Sch04]) the following condition is imposed on the models: if one *state* is indistinguishable from another, then any action name appearing for a choice in the first state also appears as an action name for a choice in the second state. It is clear right away that under this restriction, a non-uniform renaming of actions as we discussed above, may result in uncertainty relations being eliminated, and thus in a gain in knowledge. In particular, in the renamed version of example 2 above, Ann would always be able to distinguish the two states, and there would be no uncertainty left at all, which directly contradicts the requirement having to express that Ann does *not* know a uniform strategy in this situation.

4 STIT theory of agency and applications

The remaining of the course will be devoted to the role of logics of agency in multi-agent systems. Although STIT-theory was introduced by Belnap and col.'s, we will only focus on Chellas' version of STIT theory. In this section we give a summary of the main elements of the rather complex STIT semantics, BT+AC models. We warmly refer to [HB95]. (Google it!) Other relevant and more focused readings are [Tro07, BHT07b, BHT06a, BHT07a], and can be downloaded from homepages of authors. We then briefly introduce the problems related to multi-agent systems we will study in STIT models during the lecture: an embedding of Coalition Logic, and a treatment of "knowing how to play". The two remaining lectures will contain recent research. We give in this section a breviary containing key notions.

4.1 Motivation

A pinch of linguistics Already since ancient Greeks, Aristotle in *Nicomachean Ethics* by instance, philosophers have been interested in the notion of agency. It has long been a challenge to make a distinction between sentences which involve agency and those which do not. Belnap and Perloff try to uncover general principles for deciding for example whether "Ishmael sailed on board the Pequod" is agentive for Ishmael. It emphasizes a sort of causality and responsibility of an agent for the truth of a state of affairs. For Ishmael being agentive for sailing on the Pequod, there should be a prior choice of Ishmael which permitted it. (E.g. he chose deliberately to engage on the Pequod to break out of his depressive cycle.)

It is then proposed to introduce in a logical language, a binary operator reading roughly "agent i is agentive for φ ". After an analysis of several possibilities, it is decided by Belnap and Perloff that "the English verb [...] *sees to it that*, has to [their] ears at least, fewer of the obvious defects of the others, and *sees to it that* has the definite advantage of suggesting alternatives and choices" [BPX01]. It thus suggest a formal operator like $[i \textit{ stit} : \varphi]$ which reads "agent i sees to it that φ ". The STIT paraphrase thesis is as follows:

Definition 8 (STIT paraphrase thesis). *The sentence φ is agentive for agent i just in case φ may usefully paraphrased as $[i \textit{ stit} : \varphi]$. Therefore, up to an approximation, φ is agentive for i whenever $\varphi \leftrightarrow [i \textit{ stit} : \varphi]$.*

This way, deciding whether the sentence φ "Ishmael sailed on board the Pequod" is agentive for Ishmael is deciding whether it is equivalent to "Ishmael sees to it that Ishmael sailed on board the Pequod".

Agency as a modality. Origins of agency considered as a modality go back to St Anselm of Canterbury around 1100. He suggested that acting was adequately captured by what an agent brings about. He suggested that a verbal group like "killing directly" could be reformulated as "directly bringing it about that the victim is dead".

This approach gave birth to a variety of logics of action over the past fifty years that have inherited the particularity that they do not refer to the action itself but rather to its resulting state of affairs. They are logics whose main operator reads "the agent i brings it about that φ ". *We call logics of agency the logics influenced following this paradigm.*

4.2 STIT models: BT+AC

BT+AC structures are general models for logics of agency. In essence, they are reminiscent of Kutschera's proposal in [vK86]. We present here the semantics provided by Horty and Belnap in [HB95].

It is worth noting that, STIT is influenced by the observation that in a branching time framework, future-tensed statements are ambiguous to evaluate if not impossible. Suppose a moment w_0 and two

different moments w_1 and w_2 lying in the future of w_0 on two different courses of time. φ is true at w_1 and false at w_2 and everywhere before and after. What truth value should be assigned to the sentence “ φ is true in the future of w_0 ”? Indeed, φ really does lie in the future of w_1 , but what if the course of time happens to go through w_2 instead? There is a truth-value gap: in general, in branching time, a moment alone does not provide enough information to determine the truth value of a sentence about the future.

Arthur Prior [Pri67] and Richmond Thomason [Tho70, Tho84] hence proposed to evaluate future-tensed sentences with respect to a moment *and* a particular course of time running through it. This is why, as we will see, 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.

It is embedded in the branching time framework, and 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. 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. $H_w = \{h \mid h \in Hist, w \in h\}$ denotes the set of histories passing through w .

Definition 9 (index/context). *An index or context 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).*

In BT+AC 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.

In the following AGT is a non-empty set of agents and ATM is a set of atomic propositions. A $BT+AC$ -model is a tuple $\mathcal{M} = \langle W, <, Choice, v \rangle$, where:

- $\langle W, < \rangle$ is a branching time structure;
- $Choice : AGT \times W \rightarrow 2^{2^{Hist}}$ is a function mapping each agent and each moment w into a partition of H_w . The equivalence classes belonging to $Choice_i^w$ can be thought of as possible choices or actions available to i at w . Given a history $h \in H_w$, $Choice_i^w(h)$ represents the particular choice from $Choice_i^w$ containing h , or in other words, the particular action performed by i at the index w/h . We must have $Choice_i^w \neq \emptyset$ and $Q \neq \emptyset$ for every $Q \in Choice_i^w$;
- v is a valuation function $v : ATM \rightarrow 2^{W \times Hist}$.

In order to deal with group agency, Horty defines in [Hor01b, section 2.4], what he calls *group action*. Horty first introduces action selection functions s_w from AGT into 2^{H_w} satisfying the condition that for each $w \in W$ and $i \in AGT$, $s_w(i) \in Choice_i^w$. So, a selection function s_w selects a particular action for each agent at w .

Then, for a given w , $Select_w$ is the set of all selection functions s_w . For every $s_w \in Select_w$, it is assumed that

$$\bigcap_{i \in AGT} s_w(i) \neq \emptyset$$

This constraint corresponds to the assumption that the agents’ choices are independent, in the sense that agents can never be deprived of choices due to the choices made by other agents. This property is called *independence of agents* (or *independence of choices*). This constitutes a very important aspect of STIT that will be instrumental in its relationship with logics for multi-agent systems.

Using choice selection functions s_w , the $Choice$ function can be generalized to apply to groups of agents ($Choice : 2^{AGT} \times W \rightarrow 2^{2^{Hist}}$). A collective choice for a group of agents $J \subseteq AGT$ is defined as:

$$Choice_J^w = \left\{ \bigcap_{i \in J} s_w(i) \mid s_w \in Select_w \right\}$$

Again, $Choice_J^w(h) = \{h' \mid \text{there is } Q \in Choice_J^w \text{ such that } h, h' \in Q\}$.

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

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

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

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

When $\Box\varphi$ holds at one index of w then φ is said to be *settled true at w* . $\Diamond\varphi$ is defined in the usual way as $\neg\Box\neg\varphi$, and stands for historical possibility.

There are several operators in the STIT theory. The so-called *achievement* STIT was first introduced. Then it has been simplified by introducing a *deliberative* one, which is deprived of the temporal aspect, and corresponds to the previous proposition of von Kutschera [vK86, HB95]. For the purpose of this reader we only present the widely used and simpler Chellas's STIT that will be our companion for applying STIT theory to logics of multi-agent systems:³

$$\mathcal{M}, w/h \models [J]\varphi \iff \forall h' \in \text{Choice}_J^w(h), \mathcal{M}, w/h' \models \varphi$$

Intuitively $[J]\varphi$ means that group J 's current choice ensures φ , whatever the other agents do.⁴

Time operators. A quick look at the previous truth condition of the Chellas's STIT shows that by itself, it does not bring dynamics. For this reason, we will generally use time formulas as the complement of a Chellas's STIT. We have at disposition a temporal operator $\mathbf{F}\varphi$ for reasoning about future tense statements along a history:

$$\mathcal{M}, w/h \models \mathbf{F}\varphi \iff \exists w' \in h (w < w', \mathcal{M}, w'/h \models \varphi)$$

$\mathbf{F}\varphi$ reads "eventually φ holds". $[A]\mathbf{F}\varphi$ reads " A sees to it that φ eventually holds". We can combine it with historical possibility (existential quantification over histories) and form the formula $\Diamond[J]\mathbf{F}\varphi$. It reads that "it is possible that J sees to it that φ eventually holds".

4.3 Semantical comparison of CL and STIT

4.3.1 Initial settings

We have seen that the main operator of Coalition Logic $\langle\langle J \rangle\rangle\mathbf{X}\varphi$ is an operator of ability: " J is able to enforce φ at the next step whatever other agents do". This notion of 'next step' dictates us to constrain BT+AC structures so that to force the discreteness of the $<$ -ordering.

HYPOTHESIS. *Given a moment w_1 , there exists a successor moment w_2 such that $w_1 < w_2$ and there is no moment w_3 such that $w_1 < w_3 < w_2$.*

As we have discrete time, we can also define the temporal next (\mathbf{X}) operator:

$$\mathcal{M}, w/h \models \mathbf{X}\varphi \iff \exists w' \in h (w < w', \mathcal{M}, w'/h \models \varphi, \nexists w'' \in h (w < w'' < w'))$$

The next operator is not standard in STIT formalisms, but we will need it to account for a translation of the nesting of coalition modalities from CL.

Remind Section 3.4, and the fact that in Coalition Logic, when every agent has made its choice the outcome is completely determined. In order to obtain a correct embedding, we also have to enforce this in STIT-models.

HYPOTHESIS.

$$\forall w \in W, \exists w' \in W (w < w' \text{ and } \forall h \in w', \text{Choice}_{AGT}^w(h) = H_{w'})$$

³But more will be said during the course!

⁴Note the notation we use is different from the original [J cstit: φ].

Note that because $\langle W, < \rangle$ are discrete trees, the moment w' is always a next moment. From this hypothesis, the grand coalition AGT can select exactly a set of histories passing through a next moment.

4.3.2 Translating models: By the example

It can be useful to see on an example how models of CL and BT+AC models relate. We present a two-agent variant of our toy scenario of agents switching light, first in Coalition Logic and then in STIT.

Example 3. *At moment w_0 , agent i has the choice between repairing a broken lamp (ρ_i) or remaining passive (λ_i). Agent j has the vacuous choice of remaining passive : (λ_j). If i chooses not to repair, the system reaches w_1 . If i chooses to repair, the system reaches w_2 . In both w_1 and w_2 both agents can choose to toggle a light switch or not. So, agent i can choose to toggle (τ_i) or not (λ_i), and agent j can choose to toggle (τ_j) or not (λ_j).*

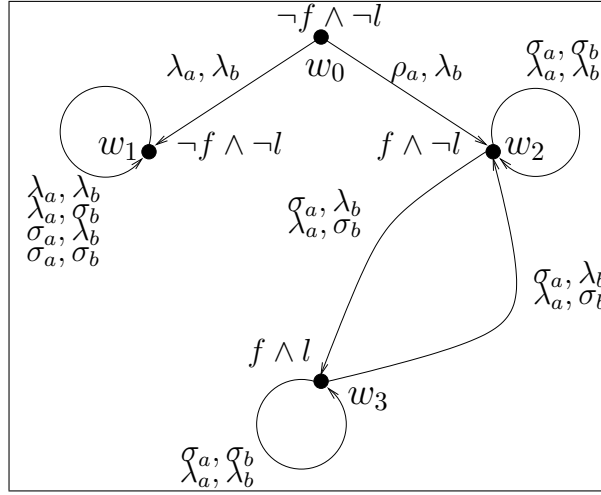


Figure 1: Game structure.

On Figure 1, the proposition f stands for “the light is functioning”, and the proposition l for “the light is on”. Now, for instance, it holds that $\mathcal{M}, w_0 \models \neg \langle i \rangle \mathbf{X} \langle j \rangle \mathbf{X} l$. So, agent i cannot ensure that agent j can ensure that the light is on. But also $\mathcal{M}, w_0 \models \langle i \rangle \mathbf{X} \langle j \rangle \mathbf{X} \neg l$. So, agent i does have a possibility (namely, choosing λ_i) that ensures that subsequently, j can avoid l . Finally, we also have that $\mathcal{M}, w_0 \models \langle i \rangle \mathbf{X} \langle j \rangle \mathbf{X} l$. That is, agent i can ensure (namely, choosing ρ_i) that the coalition $\{i, j\}$ can ensure that the light is on (namely, a choosing τ_i and j choosing λ_j or i choosing λ_i and

Figure 2 is an example of a STIT model. A feature of this model that *does not* hold for STIT models in general, is that all indexes m/h for a moment m have the same valuation of atomic propositions. For any history h through w_0 we have $\mathcal{M}, w_0/h \models \neg \diamond [i] \mathbf{X} \diamond [j] \mathbf{X} l$. Also we have $\mathcal{M}, w_0/h \models \diamond [i] \mathbf{X} \diamond [j] \mathbf{X} \neg l$ and $\mathcal{M}, w_0/h \models \diamond [i] \mathbf{X} \diamond [\{i, j\}] \mathbf{X} l$, analogous to the properties we had in the CL model.

4.3.3 A translation from CL to discrete STIT

The structural similarities between the formulas interpreted over the CL model of Figure 1 and the formulas interpreted over the STIT model of Figure 2, suggest the translation that is formalized below.

We define the translation tr from CL formulas to STIT formulae as:

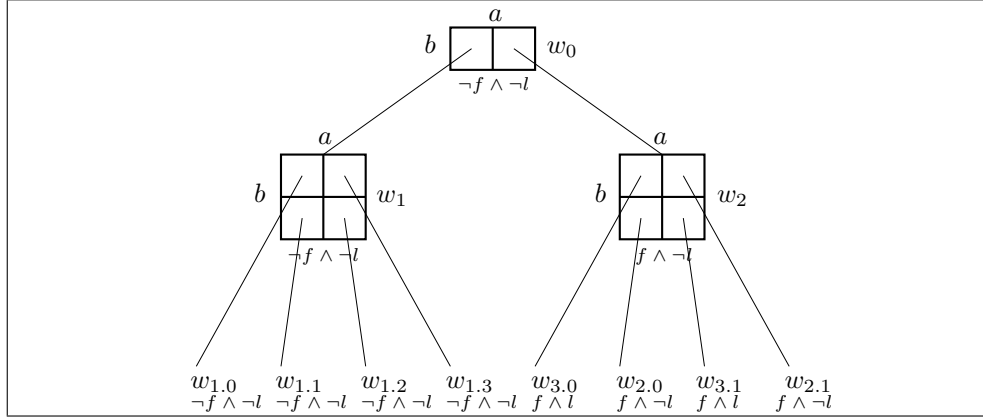


Figure 2: BT+AC model.

$$\begin{aligned}
 tr(\neg\varphi) &= \neg tr(\varphi) \\
 tr(\varphi \vee \psi) &= tr(\varphi) \vee tr(\psi) \\
 tr(\langle J \rangle \mathbf{X}\varphi) &= \diamond [J] \mathbf{X} tr(\varphi)
 \end{aligned}$$

In STIT terminology, “the coalition J is able to ensure φ ” can be paraphrased by “it is historically possible that J sees to it that next φ ”.

We refer the reader to [BHT06a] for full details.

4.4 Mathematics of STIT

4.4.1 Axiomatizing individual Chellas’s STIT (CSTIT)

While STIT has played an important role in philosophical logic since the eighties, it seems to be fair to say that its mathematical aspects have not been developed to the same extent. Most probably the reason is that STIT’s models of agency are much more complex than those existing for other modal concepts (such as say necessity, belief, or knowledge): first, the ‘seeing-to-it-that’ modalities interact (or perhaps better: must be guaranteed not to interact) because the agents’ choices are supposed to be independent; second there is another kind of modality involved, viz. the ‘master modality’ of historic necessity. As a consequence, proof systems for STIT are rather complex, too.

The language of CSTIT is built from a countably infinite set of atomic propositions ATM and a countable set of agents AGT . To simplify notation we suppose that AGT is an initial subset $\{0, 1, \dots\}$ of \mathbb{N} (possibly \mathbb{N} itself).

Formulas are built by means of the boolean connectives together with modal operators of historic necessity and of agency in the standard way.

The language of the Chellas STIT is defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid [i]\varphi \mid \Box\varphi$$

where p ranges over ATM and i ranges over AGT . This provides a standard notation for the dual constructions $\diamond\varphi$ and $\langle i \rangle\varphi$, respectively abbreviating $\neg\Box\neg\varphi$ and $\neg[i]\neg\varphi$.

Xu gave the axiomatics of Chellas’s STIT logic presented in Figure 3. The deductive system is obtained by adding standard inference rules of modus ponens and necessitation for \Box . From the latter necessitation rules for every $[i]$ follow by axiom $(\Box \rightarrow [i])$.

$(\Box \rightarrow [i])$ says that if something is settled, then every agent sees to it. This is a particularity that the Chellas STIT shares with Chellas’s agency operator used in [Che69]. Chellas’s STIT is named so

S5(\Box)	axiom schemas of S5 for \Box
S5($[i]$)	axiom schemas of S5 for every $[i]$
($\Box \rightarrow [i]$)	$\Box\varphi \rightarrow [i]\varphi$
(AIA $_k$)	$(\Diamond[0]\varphi_0 \wedge \dots \wedge \Diamond[k]\varphi_k) \rightarrow \Diamond([0]\varphi_0 \wedge \dots \wedge [k]\varphi_k)$

Figure 3: Xu’s axiomatics of Chellas’s STIT logic.

for this reason. This is a debatable property of agency that for example, renders us agentic for the Sun rising every morning.⁵

The last item is a family of *axiom schemes for independence of agents* that is parameterized by the integer k . In the particular case of two agents 0 and 1, an instance of this schemes would be:

$$\Diamond[0]\varphi_0 \wedge \Diamond[1]\varphi_1 \rightarrow \Diamond([0]\varphi_0 \wedge [1]\varphi_1)$$

Intuitively it says that if it is historically possible that agent 0 sees to it that φ_0 and that it is historically possible that agent 1 sees to it that φ_1 , then it is historically possible that *at the same time* agent 0 sees to it that φ_0 and that agent 1 sees to it that φ_1 .

This axiom is a rather complex principle. Semantically it correspond directly from the assumption of independence of agents presented on page 24. It captures the fact that whatever how agents choose, there will always be a compatible history that will be an outcome of every selected choice.

We now introduce an alternative and more standard axiomatics where we assume at least the presence of two agents. We are able to define the operator of historical necessity by exploiting the fact that because of independence of agents, an agent sees to it that a different agent sees to it that φ only if φ is settled. Formally, the following formula is a theorem of CSTIT:⁶

$$[i][j]\varphi \rightarrow \Box\varphi, i \neq j$$

This alternative axiomatic system is given axiom schemas in Figure 4 plus modus ponens and standard necessitation for $[i]$.

S5($[i]$)	axiom schemas of S5 for every $[i]$
Def(\Box)	$\Box\varphi \leftrightarrow [1][0]\varphi$
(GPerm $_k$)	$\langle l \rangle \langle m \rangle \varphi \rightarrow \langle n \rangle \bigwedge_{i \leq k, i \neq n} \langle i \rangle \varphi$ for $k \geq 0$

Figure 4: Alternative axiomatics of Chellas’s STIT logic.

(GPerm $_k$) plays the role of (AIA $_k$) and thus captures the independence of agents. It may be less intuitive, but is syntactically more suitable to study formal properties. In particular, it permits us to establish that the two-agent version of the Chellas STIT logic is nothing else than a S5 product logic presented in the first section. We refer to [BHT07a] for details.

4.4.2 Extension to groups of agents (GSTIT)

For the purpose of multi-agent systems, we need to have a system able to reason about coalitions. We present the logic GSTIT which is an extension of CSTIT to groups of agents.

⁵Variety of other STIT operators that do not have this property exist, as for example the *deliberative* STIT which can be defined as $[idstit : \varphi] \triangleq [i]\varphi \wedge \neg\Box\varphi$. An agent deliberatively sees to something only if this something is not settled.

⁶The other way round trivially holds.

Let $AGT = \{0, \dots, n-1\}$ be a finite set of $n \geq 1$ agents and ATM a countable set of atomic formulas. \mathcal{GSTIT} has the following syntax, where p ranges over elements of ATM and J ranges over the set of subsets of AGT :

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid [J]\varphi$$

The other boolean connectives are as usual defined by abbreviations, and $\langle J \rangle\varphi \stackrel{\text{def}}{=} \neg[J]\neg\varphi$. $\langle J \rangle\varphi$ roughly reads that J does not prevent φ . \bar{J} denotes the complement of J w.r.t. AGT .

We give to \mathcal{GSTIT} the axiomatization of Figure 5. We moreover have the standard inference rules

S5($[J]$)	S5 axioms for every $[J]$
(Mon)	$[J_1]\varphi \rightarrow [J_1 \cup J_2]\varphi$
Elim($[\emptyset]$)	$\langle \emptyset \rangle\varphi \rightarrow \langle J \rangle\langle \bar{J} \rangle\varphi$

Figure 5: Axiomatics of \mathcal{GSTIT} .

of modus ponens and necessitation for $[\emptyset]$. ($[\emptyset]$ plays here the role of STIT's \square .) From the latter, necessitation for every $[J]$ follows by the inclusion axiom (Mon). Note that the converse of Elim(\emptyset) can be proved from (Mon) and S5(\emptyset). Hence, we have $\vdash \langle \emptyset \rangle\varphi \leftrightarrow \langle J \rangle\langle \bar{J} \rangle\varphi$.

CSTIT is easily proved to be subsumed by \mathcal{GSTIT} . S5 nature of individual agency operators $[\{i\}]$ ($[i]$ for short) is trivial. We can derive (GPerm $_k$) in \mathcal{GSTIT} from standard S5 principles. This is left as an exercise.

Definition 10. A \mathcal{GSTIT} -model is a tuple $\mathcal{M} = (W, R, \pi)$ where:

- W is a set of worlds (alias contexts);
- R is a collection of equivalence relations R_J (one for every coalition $J \subseteq AGT$) such that:
 - $R_{J_1 \cup J_2} \subseteq R_{J_1}$
 - $R_\emptyset \subseteq R_J \circ R_{\bar{J}}$
- $\pi : W \rightarrow 2^{ATM}$ is a valuation function.

The truth conditions are:

- $\mathcal{M}, w \models p$ iff $p \in \pi(w)$
- $\mathcal{M}, w \models [J]\varphi$ iff for all $u \in R_J(w)$, $\mathcal{M}, u \models \varphi$

and as usual for the classical operators.

The fact that \mathcal{GSTIT} is determined (sound and complete) by the class of \mathcal{GSTIT} -models is immediate. All axioms are Sahlqvist-type and trivially correspond to the constraints on frames.

4.5 STIT embraces CL in the realm of normal modal logics

We capitalize on the results of the precedent section to extend propose a normal modal logic that is able to simulate both \mathcal{GSTIT} (trivially) and Coalition Logic. We call it Normal Simulation of Coalition Logic (NCL). (See [BHT07b] for details.)

4.5.1 Normal Simulation of Coalition Logic (NCL)

The language of (NCL) simply extends $\mathcal{G}STIT$'s with a 'next' temporal operator. We thus have the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathbf{X}\varphi \mid [J]\varphi$$

This way we have the discreteness we had to impose on BT+AC models in Section 4.3. We now have to be sure that the other important notion to embed CL, namely determinism, are enforced in our logic. This is the purpose of the axioms that we add to those of $\mathcal{G}STIT$ in order to obtain our axiomatics of NCL presented in Figure 6.

$\mathcal{G}STIT$	principles of $\mathcal{G}STIT$
+	
$\text{Triv}([AGT])$	$\varphi \rightarrow [AGT]\varphi$
$\mathbf{K}(\mathbf{X})$	$\mathbf{X}(\varphi \rightarrow \psi) \rightarrow (\mathbf{X}\varphi \rightarrow \mathbf{X}\psi)$
$\mathbf{D}(\mathbf{X})$	$\mathbf{X}\varphi \rightarrow \neg\mathbf{X}\neg\varphi$
$\text{Det}(\mathbf{X})$	$\neg\mathbf{X}\neg\varphi \rightarrow \mathbf{X}\varphi$

Figure 6: Axiomatics of NCL.

$\text{Triv}([AGT])$ grasps that the outcome is determined by the coalition AGT . Note that the converse of $\text{Triv}(AGT)$ is obtained by $\text{S5}(AGT)$. Hence, we have $\vdash \varphi \leftrightarrow [AGT]\varphi$. In addition, from $\mathbf{K}(\mathbf{X})$, \mathbf{X} is a normal modality. It is serial ($\mathbf{D}(\mathbf{X})$) and deterministic ($\text{Det}(\mathbf{X})$).

Definition 11. An NCL-model is a tuple $\mathcal{M} = (W, R, F_X, \pi)$ where:

- (W, R, π) is a $\mathcal{G}STIT$ -model that we constrain further such that $R_{AGT} = Id$;
- $F_X : W \rightarrow W$ is a total function;

The truth condition of \mathbf{X} is as follows:

- $\mathcal{M}, w \models \mathbf{X}\varphi$ iff $\mathcal{M}, F_X(w) \models \varphi$

Again, it is immediate that NCL is determined by the class of NCL-models by Sahlqvist theorem.

4.5.2 From CL to NCL

We give the following translation from Coalition Logic to NCL.

$$\begin{aligned} \text{tr2}(p) &= p \\ \text{tr2}(\langle J \rangle \mathbf{X}\varphi) &= \langle \emptyset \rangle [J] \mathbf{X}\text{tr}(\varphi) \end{aligned}$$

and homomorphic for the other connectives.

A group J is able to enforce φ if the nature (empty coalition) does not prevent J to see to it that next φ .

4.6 Introducing uncertainty

We have raised the problem of uniform choices in Section 3.6. We propose a solution to it in a STIT framework straightforwardly extending NCL with standard epistemic logic.

4.6.1 The Conformant STIT (ENCL)

ENCL has the following syntactic form, where p ranges over ATM , J ranges over 2^{AGT} and i over AGT :

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathbf{X}\varphi \mid [J]\varphi \mid K_i\varphi$$

The logic is obtained by adding to NCL the principles of the standard epistemic logic S5 for every individual agent i , and pictured in Figure 7.

NCL	principles of NCL
+	
S5(K_i)	S5-axioms for K_i

Figure 7: Axiomatics of ENCL.

Definition 12. An ENCL-models is tuple $\mathcal{M} = (W, R, F_X, \sim, \pi)$ where:

- (W, R, F_X, π) is a model of NCL.
- \sim is a collection of equivalence relations \sim_i (one for every agent $i \in AGT$).

Again, it is immediate that ENCL is determined by the class of ENCL-models by Sahlqvist theorem.

4.6.2 Reasoning about uniform choices

To explain how this logic solves the problem of uniform choices, we reconsider the scenarios of Section 3.6, and encode them in ENCL. The four basic properties we want to grasp are presented in Figure 8.

φ_1	<i>One of Ann's choices ensures the light will be on</i>	$\langle \emptyset \rangle [\{Ann\}] \mathbf{X}$ on
φ_2	<i>Ann knows one of her choices ensures the light will be on</i>	$K_{Ann} \langle \emptyset \rangle [\{Ann\}] \mathbf{X}$ on
φ_3	<i>Ann knows she has the power to ensure the light is on</i>	$\langle \emptyset \rangle K_{Ann} [\{Ann\}] \mathbf{X}$ on
φ_4	<i>Ann conformantly sees to it that the light is on</i>	$K_{Ann} [\{Ann\}] \mathbf{X}$ on

Figure 8: Four epistemic properties a logic of choice and knowledge should distinguish.

We show how Example 1 can be modeled as an ENCL-model. The worlds of the semantics of NCL and ENCL are here state-action pairs. The states are positions before and after execution of an action. In the picture there are six of these positions. For this example this results in eight ENCL worlds. We thus have the following ENCL-model $\mathcal{M}_1 = \langle W, R, F_X, \sim, \pi \rangle$:

- $W = \{(1, p), (1, s), (2, s), (2, p), (3, s), (4, s), (5, s), (6, s)\}$
- $R_\emptyset = \{\langle (1, p), (1, s) \rangle, \langle (2, s), (2, p) \rangle, \langle (3, s), (3, s) \rangle, \langle (4, s), (4, s) \rangle, \langle (5, s), (5, s) \rangle, \langle (6, s), (6, s) \rangle\}^*$
- $R_{Ann} = \{\langle w, w \rangle \mid w \in W\}$
- F_X is defined by $F_X((1, p)) = (3, s)$, $F_X((1, s)) = (4, s)$, $F_X((2, s)) = (5, s)$, $F_X((2, p)) = (6, s)$, $F_X((3, s)) = (3, s)$, $F_X((4, s)) = (4, s)$, $F_X((5, s)) = (5, s)$, $F_X((6, s)) = (6, s)$
- $\sim_{Ann} = \{\langle (1, p), (2, p) \rangle, \langle (1, s), (2, s) \rangle\}^*$
- π is defined by $\pi((2, p)) = \pi((2, s)) = \pi((3, s)) = \pi((5, s)) = \pi((6, s)) = \text{'on'}$, and $\pi((1, p)) = \pi((1, s)) = \pi((4, s)) = \text{'off'}$

where \star is a reflexive, symmetric and transitive closure. It is not difficult to check that \mathcal{M}_1 is a genuine ENCL-model, satisfying also all the constraints we defined for the NCL-sub-models. The reader may have noticed that the model adds detail to the example. In particular, Ann is given the choice between pushing and skipping only once, and “determinate time” is interpreted as *forever*. Of course, the model is a very simple one, with only one agent in the system: $AGT = \{Ann\}$. Ann’s actions thus coincide with system actions, and all her choices are deterministic.

It is easy to verify that in \mathcal{M}_1 the first three formulas are true in the first four possible ENCL worlds: $\mathcal{M}_1, w \models \varphi_1 \wedge \varphi_2 \wedge \varphi_3$ for all $w \in \{(1, p), (1, s), (2, s), (2, p)\}$. In particular, in the actual world $(1, p)$ the third property holds, saying that Ann has a uniform choice to ensure the light is on. In the actual world also the fourth property holds ($\mathcal{M}_1, (1, p) \models \varphi_4$), while in the two worlds where Ann skips, it does not ($\mathcal{M}_1, (1, s) \not\models \varphi_4$ and $\mathcal{M}_1, (2, s) \not\models \varphi_4$).

In turn, Example 2 can be encoded by the following ENCL-model $\mathcal{M}_2 = \langle W, R, F_X, \sim, \pi \rangle$:

- $W = \{(1, t), (1, s), (2, s), (2, t), (3, s), (4, s), (5, s), (6, s)\}$
- $R_\emptyset = \{\langle (1, t), (1, s) \rangle, \langle (2, s), (2, t) \rangle, \langle (3, s), (3, s) \rangle, \langle (4, s), (4, s) \rangle, \langle (5, s), (5, s) \rangle, \langle (6, s), (6, s) \rangle\}^\star$
- $R_{Ann} = \{\langle w, w \rangle \mid w \in W\}$
- F_X is defined by $F_X((1, t)) = (3, s)$, $F_X((1, s)) = (4, s)$, $F_X((2, s)) = (5, s)$, $F_X((2, t)) = (6, s)$, $F_X((3, s)) = (3, s)$, $F_X((4, s)) = (4, s)$, $F_X((5, s)) = (5, s)$, $F_X((6, s)) = (6, s)$
- $\sim_{Ann} = \{\langle (1, t), (2, t) \rangle, \langle (1, s), (2, s) \rangle\}^\star$
- π is defined by $\pi((2, t)) = \pi((2, s)) = \pi((3, s)) = \pi((5, s)) = \text{‘on’}$, and $\pi((1, t)) = \pi((1, s)) = \pi((4, s)) = \pi((6, s)) = \text{‘off’}$

Now, in the actual world where the light is off and Ann toggles, the light will actually be on, so the formula **Xon** holds. Yet, Ann does not conformantly see to it that the light is on, since she does not know that the light is off at the present moment. So, the fourth of the above properties does not hold: $\mathcal{M}_2, (1, t) \not\models \varphi_4$. Also, she does not have a uniform choice, and indeed the third of the above properties does not hold either: $\mathcal{M}_2, (1, t) \not\models \varphi_3$. The first and the second property do hold in the actual world, since in each state Ann indeed has an action that ensures the light is on and she knows that: But her problem is that the decision which one to take depends on the state she is in, which is something she does not know: $\mathcal{M}_2, w \models \varphi_1 \wedge \varphi_2$ for all $w \in \{(1, t), (1, s), (2, s), (2, t)\}$.

5 Intention revisited: enhancing Cohen and Levesque

The causal connection between agent and goal which is missed in C&L's logic is exactly what theories of agency such as Belnap, Horty, Chellas et col.'s 'seeing-to-it-that' STIT and Kanger, Pörn et col.'s 'bringing-it-about' provides.

Our aim here is to combine C&L's approach with Chellas' STIT operator, and argue that the resulting logic is rich enough to provide a satisfactory account of the notion of *intention to be* and of *delegation*.

5.1 A logic of agency and mental states

We have a standard possible worlds semantics for our framework, where each modal operator has logic K and is thus a normal modality. It is essentially Chellas' STIT logic of agency [BPX01].

Definition 13. *Models of intention are tuples $\langle Mom, <, Choice, B, P, v \rangle$ such that:*

- $\langle Mom, <, Choice, v \rangle$ is BT+AC model;
- B_i and P_i are accessibility relations between contexts believed and preferred by i .

We further constrain B_i and P_i such that:

- They are serial, transitive and euclidian
- B_i contains P_i (strong realism)
- if wB_iw' then $P_i(w) = P_i(w')$ (introspection)

We build upon B_i (resp. P_i) the KD45 necessity operators Bel_i (resp. $Pref_i$), defined as usual:

$$M, w/h \models Bel_i\varphi \iff \text{for all } w'/h' \in B_i(w/h), M, w'/h' \models \varphi$$

$$M, w/h \models Pref_i\varphi \iff \text{for all } w'/h' \in P_i(w/h), M, w'/h' \models \varphi$$

The temporal operators X and G are from LTL. $M, m/h \models X\varphi$ iff $M, w'/h \models \varphi$, w' being the immediate successor of w in history h . The accessibility relation for X is functional and serial, and the one for G is the reflexive and transitive closure for that of X . $F\varphi$ abbreviates $\neg G\neg\varphi$.

The following formulae on are valid:

(Stit)	S5 axioms for $[J]$;
(BoxStit)	$\Box\varphi \rightarrow [i]\varphi$;
(Monotony)	$[I]\varphi \rightarrow [J]\varphi$, for $I \subseteq J$;
(LTL)	axioms of LTL (see [Gol92]);
(Bel/Pref)	KD45 axioms for Bel_i and $Pref_i$;
(Inclusion)	$Bel_i\varphi \rightarrow Pref_i\varphi$;
(Pos. introspection)	$Pref_i\varphi \rightarrow Bel_iPref_i\varphi$;
(Neg. introspection)	$\neg Pref_i\varphi \rightarrow Bel_i\neg Pref_i\varphi$.

5.2 Intention to be

We start with a definition of **achievement goal** similar to C&L's definition. An **achievement goal** of agent i is a goal of which i does not believe it is already achieved:⁷

$$AGoal_i\varphi \stackrel{\text{def}}{=} Pref_iF\varphi \wedge \neg Bel_i\varphi.$$

Our definition of **intention to be** is:

$$Int_i\varphi \stackrel{\text{def}}{=} AGoal_i\varphi \wedge Bel_i\neg[AGT \setminus \{i\}]F\varphi.$$

Therefore according to the previous definition an agent i intends that φ iff agent i has the achievement goal that φ and believes that his intervention is needed in order to produce φ . $Bel_i\neg[AGT \setminus \{i\}]F\varphi$ is called *dependence belief*.

According to our definition of *intention to be*, an agent i cannot have the intention that it will rain or the intention that the sun will rise and so on. Indeed events such as *it rains*, *the sun rises* etc. are events φ that satisfy the following property of independence from an arbitrary agent i :

$$Indep(\varphi, i) \stackrel{\text{def}}{=} F\varphi \rightarrow [AGT \setminus \{i\}]F\varphi.$$

This means that events such as *it rains*, *the sun rises* are events whose possible future occurrence does not depend on agent i 's behavior. For instance, if it is the case that *the sun rises* then this fact is true independently from what agent i does: $Indep(SunRises, i)$. Now given an event φ (such as *the sun rises* or *it rains*) that an agent (reasonably) believes to be independent of himself, can we say that the agent intends that φ ? According to our definition this is not possible. Indeed the formula $Bel_iIndep(\varphi, i) \wedge Int_i\varphi \rightarrow \perp$ is valid in our logic. It is in this sense that we improve over C&L.

In our view the crucial aspect of the notion of *intention to be* is the fact that this is inseparable from means-end reasoning and deliberation. In order to understand what *an agent intends that* φ means, we must focus on the agent's planning activity for the achievement of φ . Our claim is the following: an agent i intends that φ only if he has decided to pursue some plans for achieving φ (viz. he intends to do something in order to achieve φ) or at least he is convinced that he must do something in order to achieve φ . We claim that this is the crucial aspect of the notion of *intention to be* and that it is nicely expressed by the *dependence belief*. Therefore if i intends that φ then either he has already decided to pursue a specific plan in order to achieve φ (viz. i intends to do something in order to achieve φ), or he is starting to build a plan in order to achieve φ .⁸ Our notion of *intention to be* is slightly different from Bratman's [Bra87]. In Bratman's theory an *intention to be* must be joined with an *intention to do*, that is, if an agent intends that p then he necessarily intends to do something in order to achieve p . Thus, according to Bratman, when an agent intends that p , he already has a plan to achieve p . In our logic, the relation between the notion of *intention to be* and the notion of *plan* is weaker than Bratman's. Our *intention to be* only needs a *dependence belief* which is the immediate precursor of an *intention to do*. Indeed, when an agent wants p to be true and believes that his intervention is needed to produce p , he is *at the beginning* of a planning process which will yield an *intention to do*.

⁷Note that we weaken C&L's negative condition $Bel_i\neg\varphi$ to $\neg Bel_i\varphi$, the reason being that $AGoal_i^{CL}F\varphi$ is inconsistent, which is contrary to intuitions.

⁸The idea that an agent builds plans in order to satisfy his goals when he believes that the achievement of what he wants depends on him, is related with a particular conception of the way instrumental intentions are generated. We adhere here to Von Wright's conception of practical inference [VW72] according to whom practical reasoning is best captured by reasoning from an end to the necessary means to that end.

6 Going fully strategic

In this section we investigate if, and how, we can generalize the theories about agency described so far to a real strategic setting. In a truly strategic setting, the powers or abilities of agents are not considered to depend on single momentary choices, but on series of choice performed one after the other. In particular we will investigate whether we can define a truly strategic notion of STIT theory.

6.1 Strategic ability: Alternating Time Temporal Logic

Alternating time temporal logic (ATL) is about abilities of agents that involve more than one choice at one particular time. Using terminology from the domain of game theory, we say that ATL is about ‘extensive form games’. Indeed, ATL can be given semantics in terms of such extensive form games.

The name ‘Alternating Time Temporal Logic’ is somewhat misleading. It is not time itself that alternates. The name stems from the original ATL semantics that was ‘turn-based’ in the sense that a strategy was thought of as a succession of choices such that agents wait for their turn and never perform a choice in parallel with another agent.

Recently ATL has emerged as a popular formalism in the agent community, for studying abilities, intentions and other Multi-agent concepts. In section 6.1.3 we will see that ATL extends both coalition logic and the branching time temporal logic CTL (for ‘Computation Tree Logic’).

6.1.1 Syntax, semantics and axiomatization of ATL

In what follows, ATM represents a set of atomic propositions, and AGT is the finite set of all agents.

Syntax Given that p ranges over ATM , and that A ranges over 2^{AGT} , the language of ATL is defined by:

$$\varphi, \psi, \dots ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\!\langle A \rangle\!\rangle \mathbf{X}\varphi \mid \langle\!\langle A \rangle\!\rangle \mathbf{G}\varphi \mid \langle\!\langle A \rangle\!\rangle \varphi \mathcal{U} \psi$$

The intended reading of $\langle\!\langle A \rangle\!\rangle \eta$, with η a linear temporal formula (branch formula), is that “group A can ensure η whatever agents in $AGT \setminus A$ do”. Again, we have adapted the original, and most used notation for the operator, which is $\langle\!\langle A \rangle\!\rangle \eta$.

Models We present models for ATL as in [AHK99], that is, in terms of Alternating Transition Systems (ATSSs) which are tuples $\mathcal{M} = \langle W, \delta, v \rangle$, where:

- W is a nonempty set of states (alias worlds, alias moments).
- $\delta : S \times AGT \rightarrow 2^{2^W}$ is a transition function mapping each moment and agent to a nonempty family of sets of possible successor moments.
- $v : ATM \rightarrow 2^W$ is a valuation function.

Each $Q \in \delta(w, a)$ may be seen as the choice by an agent of a particular action in its repertoire.

We use *lock-step synchronous* ATSSs, which means that in every state, all agents proceed simultaneously (as opposed to the particular case of *turn-based synchronous* ATSSs). The δ function is *non blocking* (agent’s actions are always compatible) and the simultaneous choice of every agent in AGT determines a *unique next state*: assuming $AGT = \{a_1, \dots, a_n\}$, for every state $w \in W$ and every set $\{Q_1, \dots, Q_n\}$ of choices $Q_i \in \delta(w, a_i)$, the intersection $Q_1 \cap \dots \cap Q_n$ is a singleton.

A *strategy for an agent a* is a mapping $f_a : W^+ \rightarrow 2^W$, such that it associates to each sequence of states $w_0 \dots w_k$ an element of $\delta(w_k, a)$.⁹ A collective strategy, for a set of agents $A \subseteq AGT$ is a

⁹it actually suffices to use mappings $f_a : W \rightarrow 2^W$ [GJ04]. However, the current definition is the customary one.

tuple $F_A = \langle f_{a_1}, f_{a_1}, \dots, f_{a_n} \rangle$ of strategies, one for each agent in AGT . The outcome of F_A from w is defined as:

$$out(w, F_A) = \{\lambda \mid \lambda = w_0 w_1 w_2 \dots, w_0 = w, \forall i \geq 0 (w_{i+1} \in \bigcap_{a \in A} f_a(w_0 \dots w_i))\}$$

A *strategy profile* is a collective strategy F_{AGT} for all agents of AGT . Analogously, a tuple $\langle Q_1, \dots, Q_n \rangle$ (one Q_i for each $i \in AGT$) is called a *choice profile*.

Semantics and axiomatization $\lambda[i]$ is the i -th position in the path λ . A formula is evaluated with respect to an ATS $\mathcal{M} = \langle W, \delta, v \rangle$ and a moment $w \in W$.

$$\begin{aligned} \mathcal{M}, w \models \langle A \rangle \mathbf{X} \varphi &\iff \exists F_A, \forall \lambda \in out(w, F_A), \mathcal{M}, \lambda[1] \models \varphi \\ \mathcal{M}, w \models \langle A \rangle \mathbf{G} \varphi &\iff \exists F_A, \forall \lambda \in out(w, F_A), \mathcal{M}, \lambda[i] \models \varphi, \forall i \geq 0 \\ \mathcal{M}, w \models \langle A \rangle \varphi \mathcal{U} \psi &\iff \exists F_A, \forall \lambda \in out(w, F_A), \\ &\exists i \geq 0 (\mathcal{M}, \lambda[i] \models \psi, \forall j \in [0, i], \mathcal{M}, \lambda[j] \models \varphi) \end{aligned}$$

Validity is defined as usual. The following complete axiomatization of ATL (as an extension of any axiomatization for propositional logic) is given in [Gvd06]. $\mathcal{M}, w \models \langle \emptyset \rangle \eta$ means that η holds irrespective of the choices made by A .

- (\perp) $\neg \langle A \rangle \mathbf{X} \perp$
- (\top) $\langle A \rangle \mathbf{X} \top$
- (N) $\neg \langle \emptyset \rangle \mathbf{X} \neg \varphi \rightarrow \langle AGT \rangle \mathbf{X} \varphi$
- (S) $\langle A_1 \rangle \mathbf{X} \varphi \wedge \langle A_2 \rangle \mathbf{X} \psi \rightarrow \langle A_1 \cup A_2 \rangle \mathbf{X} (\varphi \wedge \psi)$ if $A_1 \cap A_2 = \emptyset$
- ($FP_{\mathbf{G}}$) $\langle A \rangle \mathbf{G} \varphi \equiv \varphi \wedge \langle A \rangle \mathbf{X} \langle A \rangle \mathbf{G} \varphi$
- ($GFP_{\mathbf{G}}$) $\langle \emptyset \rangle \mathbf{G} (\theta \rightarrow (\varphi \wedge \langle A \rangle \mathbf{X} \theta)) \rightarrow \langle \emptyset \rangle \mathbf{G} (\theta \rightarrow \langle A \rangle \mathbf{G} \varphi)$
- ($FP_{\mathcal{U}}$) $\langle A \rangle \psi \mathcal{U} \varphi \equiv \varphi \vee (\psi \wedge \langle A \rangle \mathbf{X} \langle A \rangle \psi \mathcal{U} \varphi)$
- ($LFP_{\mathcal{U}}$) $\langle \emptyset \rangle \mathbf{G} ((\varphi \vee (\psi \wedge \langle A \rangle \mathbf{X} \theta)) \rightarrow \theta) \rightarrow \langle \emptyset \rangle \mathbf{G} (\langle A \rangle \psi \mathcal{U} \varphi \rightarrow \theta)$
- ($\langle A \rangle \mathbf{X}$ -Mon) from $\varphi \rightarrow \psi$ infer $\langle A \rangle \mathbf{X} \varphi \rightarrow \langle A \rangle \mathbf{X} \psi$
- ($\langle \emptyset \rangle \mathbf{G}$ -Nec) from φ infer $\langle \emptyset \rangle \mathbf{G} \varphi$

Note that the (N) axiom follows from the determinism of ‘global’ actions (actions constituted by simultaneous choices for every agent in the system): when every agent opts for a choice, the next state is fully determined, thus, if something is not settled, the coalition of all agents (AGT) can always work together to make its negation true. The axiom (S) says that two coalitions can combine their efforts to ensure a conjunction of properties if they are disjoint. Note that from (S) it follows that $\langle A_1 \rangle \varphi \wedge \langle A_2 \rangle \neg \varphi$ is not satisfiable for disjoint A_1 and A_2 . So, two disjoint coalitions cannot ensure inconsistent propositions. Axiom ($FP_{\mathbf{G}}$) characterizes the global modality as a fixpoint of the next modality, and axiom ($GFP_{\mathbf{G}}$) says that this is the greatest fixpoint. Axiom ($FP_{\mathcal{U}}$) characterizes the until operator as a (special kind of) fixpoint of the next operator, and axiom $LFP_{\mathcal{U}}$ expresses that the semantics dictates that we take the least fixpoint.

6.1.2 Game structures vs. alternating transition systems

The first paper on ATL is [AHK97]. This preliminary work is restricted to turn-based games, i.e., games where each transition is governed by a single agent. [AHK99] comes with general structures called *alternating transition systems* (ATs), where choices are expressed as sets of possible outcomes. In [AHK02] the authors change the models into *concurrent game structures* (CGSs),¹⁰ where choices are identified with explicit labels. ATs and CGSs have been proven equivalent by Goranko and Jamroga [GJ04]. Hence, defining the semantics of ATL in terms of either ATs or CGSs is a matter of convenience. However, since in its object language ATL does not refer to the labels of the choices in the game structures, we prefer the ATs as models.

6.1.3 Coalition Logic and CTL as fragments of ATL

CL (section 3) can be straightforwardly defined as a fragment of ATL. The embedding of CL in ATL should not come as a surprise given the notations we use for the central operator of both logics. The central operator of CL is $\langle\langle C \rangle\rangle X$, seen as a ‘monolith’. One of the operators of ATL is $\langle\langle C \rangle\rangle X$, but now not necessarily as one monolithic whole, since we can also have other temporal operators after the $\langle\langle C \rangle\rangle$. These other temporal operators, referring possibly to moments far in the future, are not relevant for CL, since CL only refers to the momentary choices having effect in the next system state. The embedding of CL in ATL is then just as suggested by the syntax of these operators: map the monolithic CL operator in the non-monolithic *composed* ATL operator of the same form.

To explain how CTL is a fragment of ATL, we first have to explain what CTL is. Taking ‘ p ’ to represent arbitrary elements of a given countable set of proposition symbols ATM , a well-formed formula φ of the temporal language \mathcal{L}_{CTL} is defined by:

$$\varphi, \psi, \dots := p \mid \neg\varphi \mid \varphi \wedge \psi \mid E(\varphi U^{ee}\psi) \mid A(\varphi U^{ee}\psi)$$

where φ, ψ represent arbitrary well-formed formulas. The version of until we use does not require φ to hold in the starting state or ending state of the path on which it is evaluated. This enables us to define the next operation and other operators as abbreviations.

$$\begin{array}{ll} EX\varphi \equiv_{def} E(\perp U^{ee}\varphi) & AX\varphi \equiv_{def} \neg EX\neg\varphi \\ EF\varphi \equiv_{def} E(\top U^{ee}\varphi) & AG\varphi \equiv_{def} \neg EF\neg\varphi \\ AF\varphi \equiv_{def} A(\top U^{ee}\varphi) & EG\varphi \equiv_{def} \neg AF\neg\varphi \end{array}$$

The CTL-operators have the following informal meanings:

$E(\varphi U^{ee}\psi)$:	there is a possible future course of action after which ψ will hold, while φ holds from the next moment until then
$A(\varphi U^{ee}\psi)$:	for all possible future courses of action eventually ψ will hold, while φ holds from the next moment until then
$EX\varphi$:	there is an atomic action after which φ will hold
$AX\varphi$:	after application of any atomic action φ will hold
$EF\varphi$:	there is a possible future course of action after which φ will hold
$AG\varphi$:	for all possible future courses of action φ will be preserved
$AF\varphi$:	for all possible future courses of action eventually φ will hold
$EG\varphi$:	there is a possible future course of action that preserves φ

A CTL structure is a tuple $M = \langle W, \mathcal{R}, V \rangle$ where W is a set of moments, $\mathcal{R} \subseteq W \times W$ is a serial relation, and V an interpretation function that assigns a particular set of moments to each primitive proposition: $V(p)$ contains all those situations in which p holds. The formal semantics of the CTL-operators can be given in terms of ‘infinite paths’ through these CTL structures. Since R is serial, any

¹⁰An alternative name from the literature is ‘multi-player game model’, abbreviated ‘MGM’.

finite path can always be extended to an infinite one. We could also have taken infinite trees for the models. The semantics over trees is where CTL got its name from.

An infinite path in a CTL-structure M is a sequence $\sigma = m_0, m_1, m_2, \dots$ such that for every $i \geq 0$, m_i is an element of W and $m_i \mathcal{R} m_{i+1}$. We say that an infinite path starts at m iff $m_0 = m$. If $\sigma = m_0, m_1, m_2, \dots$ is an infinite path in M , then we denote m_i by σ^i ($i \geq 0$).

Let M be a CTL-structure, m a situation, and σ an infinite path. The semantic relation \models for CTL is then defined as follows:

- $M, m \models p$ iff $m \in V(p)$ and p is a primitive proposition
- $M, m \models \alpha \wedge \beta$ iff $m \models \alpha$ and $m \models \beta$
- $M, m \models \neg\alpha$ iff $m \models \alpha$ does not hold
- $M, m \models E\alpha$ iff \exists infinite path σ in M starting at m s.t. $M, \sigma \models \alpha$
- $M, m \models A\alpha$ iff \forall infinite path σ in M starting at m , it holds that $M, \sigma \models \alpha$
- $M, \sigma \models \alpha U^{ee} \beta$ iff there is at least one $i \geq 0$ such that $M, \sigma^i \models \beta$ and for all j with $(0 < j < i)$ it holds that $M, \sigma^j \models \alpha$

Now it is quite straightforward in what sense ATL generalized CTL. CTL can be seen as a fragment of ATL using the following definitions.

$$\begin{aligned} E(\varphi U^{ee} \psi) &\equiv_{def} \langle \langle AGT \rangle \rangle (\varphi U^{ee} \psi) \\ A(\varphi U^{ee} \psi) &\equiv_{def} \langle \langle \emptyset \rangle \rangle (\varphi U^{ee} \psi) \end{aligned}$$

Of course, strictly speaking, this does not give an embedding of the ATL language given in section 6.1.1, but this can easily be remedied. We can reformulate the semantics of ATL entirely in terms of the operator $\langle \langle J \rangle \rangle (\varphi U^{ee} \psi)$. Actually, we will do so in section 6.3.

6.2 Embedding ATL into strategic STIT ability

In chapter 6 of his 2001 book [Hor01b] Horty gives a semantic account of a logic of strategic STIT ability (our terminology). This logic comes really close to ATL (and even ATL*), and in this section we explain the relation. The syntax of Horty's logic of strategic STIT ability is defined as follows.

Syntax Given that p ranges over ATM , and that A ranges over 2^{AGT} , a language of strategic STIT ability is defined by:

$$\varphi, \psi, \dots ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid \mathbf{X}\varphi \mid \mathbf{G}\varphi \mid \varphi \mathcal{U} \psi \mid \Diamond_s[A \textit{scstit} : \varphi]$$

First we have to explain why we call the logic defined relative to the above syntax a logic of 'strategic STIT *ability*' in stead of a logic of 'strategic STIT'. The intuitive reading of $\Diamond_s[A \textit{scstit} : \varphi]$ is "it is strategically possible that agents A see to it that φ ". The operator $\Diamond_s[A \textit{scstit} : \varphi]$, suggested by Horty [Hor01b, p.152], is thus a monolithic ('fused', as Horty calls it) operator that is 'built' from an operator for strategic possibility ($\Diamond_s\varphi$) and a strategic version of Chellas' STIT operator ($[A]\varphi$). However, in Horty's work these separate operators are not given a formal semantics individually; the operators are syntactically forced to occur only in combination (in section 6.3 we solve this problem). Yet, to understand the semantics of the fused operator, below we discuss the intended semantics of the individual operators.

Like in section 4 the semantics 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: 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 $w_2 < w_1$.

A maximal set of linearly ordered moments from W is a *history*. Thus, $w \in h$ denotes that 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).

To enable a comparison with ATL we make the following assumption:

Assumption 1 (countably infiniteness). *Every history is isomorphic to the set of natural numbers.*

By assuming that histories are countably infinite sets of moments we will be able to reason about temporal properties as in LTL.

A *STIT model* is a tuple $\mathcal{M} = \langle W, Choice, <, v \rangle$, where:

- $\langle W, < \rangle$ is a branching-time structure.
- $Choice : AGT \times W \rightarrow 2^{2^{Hist}}$ 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 agent a at w . Given a history $h \in H_w$, $Choice_a^w(h)$ represents the particular 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 $Q \in Choice_a^w$.
- v is valuation function $v : ATM \rightarrow 2^{W \times Hist}$.

At index w/h we shall call w the *current moment* and $Choice_a^w(h)$ the *current choice/action*. In order to deal with group agency, Horty defines in [Hor01b, section 2.4], the notion of collective choice. Horty first introduces action selection functions s_w from AGT into 2^{H_w} satisfying the condition that for each $w \in W$ and $a \in AGT$, $s_w(a) \in Choice_a^w$. So, a selection function s_w selects a particular action for each agent at w .

Then, for a given w , $Select_w$ is the set of all selection functions s_w . For every $s_w \in Select_w$, it is assumed that $\bigcap_{a \in AGT} s_w(a) \neq \emptyset$. This constraint corresponds to the assumption that the agents' choices are independent, in the sense that agents can never be deprived of choices due to the choices made by other agents.

Moreover, in order to match ATL, we make the following assumption stating that the intersection of choices of agents in AGT must exactly be the set of histories passing through some immediate next moment:

Assumption 2 (determinism).

$$\forall w \in W, \exists w' \in W (w < w' \text{ and } \bigcap_{a \in AGT} s_w(a) = H_{w'})$$

Note that because STIT frames are trees, the moment w' is always a next moment.

Using choice selection functions s_w , the *Choice* function can be generalized to apply to groups of agents ($Choice : 2^{AGT} \times W \rightarrow 2^{2^{Hist}}$). A collective choice for a group of agents $A \subseteq AGT$ is defined as:

$$Choice_A^w = \left\{ \bigcap_{a \in A} s_w(a) \mid s_w \in Select_w \right\}$$

Again, $Choice_A^w(h) = \{h' \mid \text{there is } Q \in Choice_A^w \text{ such that } h, h' \in Q\}$.

Strategies [Hor01b, BPX01] introduce strategies into STIT theory: a *strategy* for an agent a is a partial function σ on W such that $\sigma(w) \in Choice_a^w$ for each moment w from $Dom(\sigma)$, the domain of σ . In STIT theory it is assumed that σ may be a partial function. The reason is that there is no need to account for choices at states an agent never arrives at by following σ . In [BPX01, p.350] it says ‘‘A strategy need not tell us what to do at moments that the strategy itself forbids’’. This contrasts with

ATL, where it is implicitly assumed that strategies are total. But, as the present comparison between both systems reveals, for the basic ATL modalities this is not at all necessary.¹¹

As we can see in the definition of the $[_]_$ operator, an agent's choice restricts the set of possible futures, in particular it restricts the histories to those corresponding with the choice being made. We expect a strategy to be a generalization of this, in particular, we want a strategy to restrict the possible histories to those corresponding to a series of choices being made at successive moments.

A strategy σ *admits* a history h if and only if (i) $Dom(\sigma) \cap h \neq \emptyset$ and (ii) for each $w \in Dom(\sigma) \cap h$ we have $h \in \sigma(w)$. The set of all histories admitted by a strategy σ is denoted $Adh(\sigma)$.

We will often use the notation σ_a , to name a particular strategy of an agent a .

A collective strategy for $A \subseteq AGT$ is a tuple $\sigma_A = \langle \sigma_a \rangle_{a \in A}$, and $Adh(\sigma_A) = \bigcap_{a \in A} Adh(\sigma_a)$.

Horty [Hor01b] also proposes strategies with a limited scope. To this end, he introduces the notion of *field* which is a $<$ -backward closed subset M of $Tree_w = \{w' \mid w < w' \text{ or } w = w'\}$. With $Adm(\sigma) = \{w \mid w \in h, h \in Adh(\sigma)\}$, a strategy is properly formed in the field M if it is *complete* in M ($Adm(\sigma) \cap M \subseteq Dom(\sigma)$) and *irredundant* ($Dom(\sigma) \subseteq Adm(\sigma)$). Thus, an *ability* operator should be evaluated with respect to a field.

Here we do not need such a refinement. Therefore, for any strategy at a moment w we will always consider the field to be the complete set $Tree_w$, that is, the backward-closed sub-tree having w as root. For evaluation of formulas in the strategic setting we will use the same models and indexes as for the non-strategic setting.¹²

As discussed in [Hor01b], global effectivity by means of a strategy differs from local effectivity induced by a unique (possibly collective) choice. Available choices at a moment form a partition of that moment: one history lies in one and only one choice. But, the sets of admitted histories of the strategies available at a given moment do *not* necessarily partition that moment. One history can lie in the sets of admitted histories of two different strategies. Therefore, since a history alone does not tell us which strategy we have to consider, we cannot evaluate global effectivity as we have done for local effectivity (the $[_]_$ operator). However, those semantic difficulties are outside the scope of this paper. We refer the reader to [Hor01b, Section 7.2.1] and to [BHT06b], where we propose a solution to this problem in the ATL-setting.

Horty points out that we can return to a natural evaluation by using an operator quantifying over strategies. In particular, we can define the *fused* operator for long term strategic ability of groups of agents as follows:

Semantics $\mathcal{M}, w/h \models \Diamond_s[A \textit{ scstit} : \varphi] \iff \exists \sigma \in Strategy_A^w \text{ s.t. } \forall h' \in Adh(\sigma), \mathcal{M}, w/h' \models \varphi$

where $Strategy_A^w = \{\sigma \mid Dom(\sigma) = Tree_w\}$, is the set of strategies open to A at moment w .¹³

Intended readings for $\Diamond_s[A \textit{ scstit} : \varphi]$ are: “it is strategically possible that agents A see to it that φ ”, or “ A has the ability to guarantee the truth of φ by carrying out an available strategy”. Horty uses a slightly different syntax and writes this fused operator as $\Diamond[A \textit{ scstit} : \varphi]$. We use the s -subscript for the diamond to emphasize that it does not reflect *historical* possibility (written without the s -subscript as $\Diamond\varphi$) but *strategic* possibility. For enlightenment, we mention the connections of this operator with Chellas' STIT operator and the historical necessity operator.

¹¹However, if we extend ATL with strategic STIT operators, as we did in [BHT06b], totality of strategy functions with respect to the domain of states is indeed necessary.

¹²It is easy to see that actually histories are not needed to evaluate the strategic ability operator. Horty calls this moment-determinateness of the fused operator. We nevertheless keep the histories for uniformity purposes.

¹³In the original definition, a set of strategies is denoted $Strategy_A^M$, where M is a field having w as root. Since we have assumed that M is always $Tree_w$, our notation $Strategy_A^w$ suffices.

The strategic ability operator $\diamond_s[A \text{ scstit} : \varphi]$ can be seen to be stronger than the local ability operator $\diamond[_]_{_}$. In particular, it holds that:

$$\models_{STIT} \diamond[A]\varphi \rightarrow \diamond_s[A \text{ scstit} : \varphi].$$

This property ensures that the translation we propose below for ATL, embeds the translation we did for CL in section 3.4.

Now we are ready to give the translation of ATL into Horty's strategic STIT ability.

$$\begin{aligned} tr(p) &= \Box p, \text{ for } p \in ATM \\ tr(\neg\varphi) &= \neg tr(\varphi) \\ tr(\varphi \vee \psi) &= tr(\varphi) \vee tr(\psi) \\ tr(\langle A \rangle \mathbf{X}\varphi) &= \diamond_s[A \text{ scstit} : \mathbf{X}tr(\varphi)] \\ tr(\langle A \rangle \mathbf{G}\varphi) &= \diamond_s[A \text{ scstit} : \mathbf{G}tr(\varphi)] \\ tr(\langle A \rangle \varphi \mathcal{U} \psi) &= \diamond_s[A \text{ scstit} : tr(\varphi) \mathcal{U} tr(\psi)] \end{aligned}$$

Our strategy for proving that this determines a correct embedding, is to show that (1) If φ is ATL-satisfiable then $tr(\varphi)$ is STIT-satisfiable, and that (2) if $\models_{ATL} \varphi$ then $\models_{STIT} tr(\varphi)$. For this second direction of the proof, we use the ATL axiomatization of [GvD06], and prove that translation of the axioms are valid, and that the translated inference rules preserve validity. For the first direction we use a model construction. Here we only briefly discuss how these models look like.

Given a tree-like choice partitioned ATS $\mathcal{M}_{ATL} = \langle W_{ATL}, \delta, v_{ATL} \rangle$ we associate to it a STIT model $\mathcal{M}_{STIT} = \langle W_{STIT}, Choice, <, v_{STIT} \rangle$, as follows:

- $W_{STIT} = W_{ATL}$
- $w < u \iff \exists u_1, \dots, u_n (u_1 = w, u_n = u, \forall i < n (\exists a \in AGT, Q_a \in \delta(u_i, a), u_{i+1} \in Q_a))$
- $Choice_a^w = \{\{h | Q_a \cap h \neq \emptyset\} | Q_a \in \delta(w, a)\}$ for all a and m
- $\forall h \in H_w, v_{STIT}(w/h) = v_{ATL}(w)$

It is clear that the tree property is instrumental for $\langle W_{STIT}, < \rangle$ being a tree. We inherit the branching-time structure of STIT directly from the tree structure of the ATS. Furthermore, the condition concerning partitions underlying choice partitioned ATSs prevents that two choices of the same agent have a non-empty intersection, and therefore every $Choice_a^w$ is a partition of H_w . If intersections would possibly be non-empty, we could not have constructed the $Choice$ function as we did: the same history could have been in two different sets of $Choice_a^w$.

6.3 Strategic STIT

In the previous (sub)section we saw that under some extra assumptions, we could map ATL in a version of strategic STIT ability as defined by Horty. However it would be better if we would have truly strategic notion of STIT itself (instead of STIT-ability). In this section we present a semantics for a strategic version of STIT, combined with modal operators for quantification over strategies. This logic, called ATL-STIT here, embeds ATL. Again we use a non-standard, but concise and intuitive syntax and semantics.

6.3.1 Core Syntax, Abbreviations and Intended Meanings

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

$$\begin{aligned} \varphi, \psi, \dots &:= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \diamond_A \varphi \mid \Box_A \varphi \mid [A]\eta \mid \langle A \rangle \eta \\ \eta, \theta, \dots &:= \varphi U^{ee} \psi \end{aligned}$$

where φ, ψ, \dots represent arbitrary well-formed formulas, η, θ, \dots represent temporal path formulas, the p are elements from an infinite set of propositional symbols ATM , and A is a subset of a finite set of agent names AGT (we define $\bar{A} \equiv_{def} AGT \setminus A$). 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, as in our exposition of CTL in section 6.1.3. 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 $\langle\langle A \rangle\rangle \eta$, while we write it as $\langle [A] \rangle \eta$), 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 \bar{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 \bar{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.

$$\begin{array}{ll}
[A]X\varphi \equiv_{def} [A](\perp U^{ee} \varphi) & \langle A \rangle X\varphi \equiv_{def} \langle A \rangle(\perp U^{ee} \varphi) \\
[A]F\varphi \equiv_{def} [A](\top U^{ee} \varphi) & \langle A \rangle F\varphi \equiv_{def} \langle A \rangle(\top U^{ee} \varphi) \\
[A]G\varphi \equiv_{def} \neg \langle A \rangle F\neg\varphi & \langle A \rangle G\varphi \equiv_{def} \neg [A]F\neg\varphi \\
[A](\varphi U^e \psi) \equiv_{def} [A](\varphi U^{ee}(\varphi \wedge \psi)) & \langle A \rangle(\varphi U^e \psi) \equiv_{def} \langle A \rangle(\varphi U^{ee}(\varphi \wedge \psi)) \\
[A](\varphi U_w^e \psi) \equiv_{def} \neg \langle A \rangle(\neg \psi U^e \neg \varphi) & \langle A \rangle(\varphi U_w^e \psi) \equiv_{def} \neg [A](\neg \psi U^e \neg \varphi)
\end{array}$$

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
- $[A](\varphi U_w^e \psi)$: 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 language, 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.

6.3.2 Model theoretic semantics

We use alternating transition systems (ATSS) for the semantics. 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 all 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.

An ATSS $\mathcal{M} = (S, \mathcal{C}, \pi)$, consists of a non-empty set S of states, a total function $\mathcal{C} : AGT \times S \mapsto 2^{2^S}$ 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 $RX(s) = \{ \bigcap_{a \in AGT} Ch_a \mid Ch_a \in \mathcal{C}(a, s) \}$ 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 \mathcal{C} it follows that the choices for each individual agent at a certain moment in time partition 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.

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

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 (though ATL* is).

Strategy functions α_a for individual agents a are straightforwardly combined to system strategy functions $\alpha_{AGT} : S \times AGT \mapsto 2^S$ for the full set of agents AGT . Then $\alpha_{AGT}(s, a)$ yields the choice of agent a in state s determined by the system strategy α_{AGT} . However, central to our semantics will be *partial* strategy functions $\alpha_A : S \times AGT \mapsto 2^S$, where $A \subseteq AGT$. These functions are partial in the sense that no choices are defined for the agents \bar{A} . For $B \subseteq A$ we use the notation $\alpha_A \upharpoonright_B$ 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 $\alpha_A \upharpoonright_A = \alpha_A$). Furthermore, for $A \cap B = \emptyset$, we use $\alpha_A \mid \beta_B$ to denote the minimal joined partial strategy function build from α_A and β_B such that $(\alpha_A \mid \beta_B) \upharpoonright_A = \alpha_A$ and $(\alpha_A \mid \beta_B) \upharpoonright_B = \beta_B$.

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.

Given a system strategy α_{AGT} , we define the follow up function $F_{\alpha_{AGT}} : S \mapsto S$ as the intersection of all choices for individual agents, that is, $F_{\alpha_{AGT}}(s) = \bigcap_{a \in AGT} \alpha_{AGT}(s, a)$.

Then, by $(F_{\alpha_{AGT}})^n(s)$ we denote the unique state that results from state s by taking n steps of the system strategy α_{AGT} .

Now we are ready to define the formal semantics of the language $\mathcal{L}_{ATL-STIT}$. 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 (see also section 1.5 on products), we have modal operators interpreted on individual dimensions only: the strategy quantification operators $\diamond_A\varphi$ and $\square_A\varphi$ are interpreted on the dimension of strategies, relative to a *fixed* moment, and the temporal STIT operators $[A]\varphi U^{ee}\psi$ and $\langle A\rangle\varphi U^{ee}\psi$ 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.

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 $\mathcal{M} = (S, \mathcal{C}, \pi)$ is defined as:

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

$$\begin{array}{ll}
\mathcal{M}, \alpha_A, s \models \langle B\rangle\varphi U^{ee}\psi & \Leftrightarrow \exists \beta_{\overline{A \cap B}} \text{ and } \exists n > 0 \text{ such that} \\
& \quad (1) \mathcal{M}, \alpha_A, (F_{\alpha_{AGT}})^n(s) \models \psi \text{ and} \\
& \quad (2) \forall i \text{ with } 0 < i < n \text{ we have } \mathcal{M}, \alpha_A, (F_{\alpha_{AGT}})^i(s) \models \varphi \\
& \quad \text{where } \alpha_{AGT} \text{ is defined as: } \alpha_{AGT} = \alpha_A \upharpoonright_{A \cap B} \upharpoonright_{\beta_{\overline{A \cap B}}}
\end{array}$$

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 $\mathcal{L}_{\text{ATL-STIT}}$ 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 the STIT formalisms in section 4 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.

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

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 [Hor01a] 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.

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

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 $\langle A \rangle \eta \equiv_{def} \diamond_A [A] \eta$ 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.

We will only discuss a few validities of ATL-STIT. We do not give an axiomatization. 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 accessibility 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.

6.4 Epistemic strategic 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 strategy / state pairs. The resulting fine-grained epistemic structures enable us to reconsider the problem of so called ‘uniform strategies’. The difference with section 3.6 is that the solution given there, does not work for the fully strategic case. Here we show how to approach that problem.

6.4.1 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 .

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

$$\begin{aligned} \varphi, \psi, \dots &:= p \mid \neg \varphi \mid \varphi \wedge \psi \mid K_a \varphi \mid E_A \varphi \mid D_A \varphi \mid C_A \varphi \mid \diamond_A \varphi \mid \Box_A \varphi \mid [A] \eta \mid \langle A \rangle \eta \\ \eta, \theta, \dots &:= \varphi U^{ee} \psi \end{aligned}$$

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 [JH04], 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.

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 6.3.2 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.

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.

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

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.

6.4.2 The problem of uniform strategies

The most discussed problem for epistemic additions to ATL discussed in the literature (ATEL [HW03]), is the problem of so called ‘uniform strategies’. We briefly recall the problem by means of the cards example from [JH04] (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 $K_a \langle a \rangle \text{win}$.

This holds since the agent cannot distinguish the state where he has the winning card from the state where he has the losing 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 [HW03] 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 [NM44]. 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 [HT06] 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.

References

- [AHK97] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pages 100–109. IEEE Computer Society Press, 1997.
- [AHK99] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. In *Compositionality: The Significant Difference*, Lecture Notes in Computer Science 1536, pages 23–60. Springer, 1999.
- [AHK02] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49:672–713, 2002.
- [AS87] B. Alpern and F.B. Schneider. Recognizing safety and liveness. *Distributed Computing*, 2:117–126, 1987.
- [BHT06a] Jan Broersen, Andreas Herzig, and Nicolas Troquard. From Coalition Logic to STIT . In Wiebe van der Hoek, Alessio Lomuscio, Erik de Vink, and Mike Wooldridge, editors, *Third International Workshop on Logic and Communication in Multi-Agent Systems (LCMAS 2005)* , Edinburgh, Scotland, UK, volume 157:4 of *Electronic Notes in Theoretical Computer Science*, pages 23–35. Elsevier, 2006.
- [BHT06b] J.M. Broersen, A. Herzig, and N. Troquard. A STIT-extension of ATL. In Michael Fisher, editor, *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA '06)*, volume 4160 of *Lecture Notes in Artificial Intelligence*, pages 69–81. Springer Verlag, 2006.
- [BHT07a] P. Balbiani, A. Herzig, and N. Troquard. Alternative axiomatics and complexity of deliberative STIT theories, 2007. arXiv:0704.3238v1. Submitted.
- [BHT07b] Jan Broersen, Andreas Herzig, and Nicolas Troquard. Normal simulation of coalition logic and an epistemic extension. In *Proceedings of TARK 2007*, Brussels, Belgium, 2007. ACM DL.
- [BPX01] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford, 2001.
- [Bra87] M. E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- [BRV01] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 2001.
- [Che69] Brian Chellas. *The Logical Form of Imperatives*. PhD thesis, Philosophy Department, Stanford University, 1969.
- [Chi63] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.
- [CL90] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence Journal*, 42(2–3):213–261, 1990.
- [GJ04] V. Goranko and W.J. Jamroga. Comparing semantics of logics for multi-agent systems. *Synthese*, 139(2):241–280, 2004.
- [Gol92] R. Goldblatt. *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California, 1992.

- [GvD06] Valentin Goranko and Govert van Drimmelen. Decidability and complete axiomatization of the alternating-time temporal logic. *Theoretical Computer Science*, 353(1-3):93–117, 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.
- [HKT00] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. The MIT Press, 2000.
- [HL02] Andreas Herzig and Dominique Longin. Sensing and revision in a modal logic of belief and action. In Frank van Harmelen, editor, *Proc. ECAI2002*, pages 307–311. IOS Press, 2002.
- [HL04] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning(KR2004)*, pages 527–535. AAAI Press, 2004.
- [Hor01a] J.F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.
- [Hor01b] John F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.
- [HT06] Andreas Herzig and Nicolas Troquard. Knowing How to Play: Uniform Choices in Logics of Agency. In Gerhard Weiss and Peter Stone, editors, *5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan*, pages 209–216. ACM Press, 8-12 May 2006.
- [HW03] W. van der Hoek and M. Wooldridge. Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75(1):125–157, 2003.
- [JH04] W.J. Jamroga and W. van der Hoek. Agents that know how to play. *Fundamenta Informaticae*, 63(2), 2004.
- [LHC06] Emiliano Lorini, Andreas Herzig, and Cristiano Castelfranchi. Introducing attempt in a modal logic of intentional action. In Michael Fisher and Wiebe van der Hoek, editors, *Proc. 10th Eur. Conf. on Logics in Artificial Intelligence (JELIA06)*, volume 4160 of *LNAI*, pages 1–13, Liverpool, 13-15 September 2006. Springer-Verlag.
- [MHL99] J.-J.Ch Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 1999.
- [Moo85] Robert C. Moore. A formal theory of knowledge and action. In J.R. Hobbs and R.C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex, Norwood, NJ, 1985.
- [NM44] J. von Neumann and O. Morgenstern. *Theory of games and economic behaviour*. Princeton University Press, 1944.
- [Pau02] Marc Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
- [Pra76] V.R. Pratt. Semantical considerations on Floyd-Hoare logic. In *Proceedings 17th IEEE Symposium on the Foundations of Computer Science*, pages 109–121. IEEE Computer Society Press, 1976.
- [Pri67] A.N. Prior. *Past, Present, and Future*. Clarendon Press, 1967.
- [RG95] A.S. Rao and M.P. Georgeff. Formal models and decision procedures for multi-agent systems. Technical Report Technical Note 61, Melbourne, Australia, 1995.

- [Sad00] M. D. Sadek. Dialogue acts are rational plans. In M.M. Taylor, F. Néel, and D.G. Bouwhuis, editors, *The structure of multimodal dialogue*, pages 167–188, Philadelphia/Amsterdam, 2000. John Benjamins publishing company. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991.
- [Sch04] P. Y. Schobbens. Alternating-time logic with imperfect recall. *Electronic Notes in Theoretical Computer Science*, 85(2), 2004.
- [Sea83] J. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, 1983.
- [SL93] Richard Scherl and Hector J. Levesque. The frame problem and knowledge producing actions. In *Proc. Nat. Conf. on AI (AAAI'93)*, pages 689–695. AAAI Press, 1993.
- [SL03] Richard Scherl and Hector J. Levesque. The frame problem and knowledge producing actions. *Artificial Intelligence*, 144(1-2), 2003.
- [SPLL00] S. Shapiro, M. Pagnucco, Y. Lespérance, and H. J. Levesque. Iterated belief change in the situation calculus. In *Proc. KR2000*, pages 527–538, 2000.
- [Tar41] A. Tarski. On the calculus of relations. *Journal of Symbolic Logic*, 6:73–89, 1941.
- [Tho70] Richmond Thomason. Indeterminist time and truth-value gaps. *Theoria*, 36:264–81, 1970.
- [Tho84] Richmond H. Thomason. Combinations of tense and modality. In Dov Gabbay and Franz Guenther, editors, *Handbook of Philosophical Logic: Extensions of Classical Logic*, pages 135–165. Reidel, 1984.
- [Tho00] R. H. Thomason. Desires and defaults: A framework for planning with inferred goals. In *Proc. of the Seventh International Conference Knowledge Representation*. Morgan Kaufmann Publishers, 2000.
- [Tro07] Nicolas Troquard. *Independent agents in branching time*. PhD thesis, Université Paul Sabatier, Toulouse, 2007.
- [vK86] Franz von Kutschera. Bewirken. *Erkenntnis*, 24(3):253–281, 1986.
- [VW72] G. H. Von Wright. On so-called practical inference. *The Philosophical Review*, 15:39–53, 1972.
- [Woo02] Michael Wooldridge. *Introduction to MultiAgent Systems*. John Wiley and Sons, 2002.
- [Wri51] G.H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.