Fixpoint Approximation of Strategic Abilities under Imperfect Information

Paper no. 2444

Abstract

Model checking of strategic ability under imperfect information is known to be hard. In this paper, we propose translations of $\mathbf{ATL}_{\mathrm{ir}}$ formulae that provide lower and upper bounds for their truth values, and are cheaper to verify than the original specifications. Most interestingly, the lower approximation is provided by a fixpoint expression that uses a nonstandard variant of the next-step ability operator. We show the correctness of the translations, establish their computational complexity, and validate the approach by experiments with a scalable scenario of Bridge play.

Introduction

There is a growing number of works that study *strategic logics*, in particular syntactic and semantic variants of alternating-time temporal logic **ATL** for agents with imperfect information (Ågotnes et al. 2015). The contributions are mainly theoretical, and include results concerning the conceptual soundness of a given semantics of ability (Schobbens 2004; Ågotnes 2004; Jamroga and van der Hoek 2004; Dima, Enea, and Guelev 2010; Guelev and Dima 2012), meta-logical properties (Guelev, Dima, and Enea 2011; Bulling and Jamroga 2014), and the complexity of model checking (Schobbens 2004; Jamroga and Dix 2006; Guelev, Dima, and Enea 2011; van der Hoek, Lomuscio, and Wooldridge 2006; Dima and Tiplea 2011). However, there is relatively little research on the actual *use* of the logics, in particular on practical algorithms for verification in scenarios where agents have a limited view of the world.

This is somewhat easy to understand, since model checking of **ATL** variants with imperfect information has been proved $\mathbf{\Delta_2^P}$- to **PSPACE**-complete for agents playing memoryless strategies and undecidable for agents with perfect recall of the past. Moreover, the imperfect information semantics of **ATL** does not admit fixpoint equivalences (Bulling and Jamroga 2014), which makes incremental synthesis of strategies impossible, or at least difficult to achieve. Some early attempts at verification of imperfect information strategies made their way into the MCMAS model-checker (Lomuscio and Raimondi 2006; Raimondi 2006; Lomuscio, Qu, and Raimondi 2015), but the issue was never at the heart of the tool. More dedicated attempts began to emerge only recently (Pilecki, Bednarczyk, and Jamroga 2014; Busard et al. 2014; Huang and van der Meyden 2014; Busard et al. 2015). Up until now, experimental results confirm that the initial intuition was right: model checking strategic modalities for imperfect information is hard, and dealing with it requires innovative verification techniques.

In this paper, we propose that in some instances, instead of exact model checking, it suffices to provide an upper and/or lower bound for the output. The intuition for the upper bound is straightforward: instead of checking existence of imperfect information strategy, we can look for a perfect information strategy that obtains the same goal. If the latter is false, the former must be false too. Finding a reasonable lower bound is nontrivial, but we construct one by means of a fixpoint expression in alternating epistemic mu-calculus. We begin by showing that the straightforward fixpoint approach does not work. Then, we propose how it can be modified to obtain guaranteed lower bounds. To this end, we alter the next-step operator in such a way that traversing the appropriate epistemic neighborhood is seen as an atomic activity. We show the correctness of the translations, establish their computational complexity, and validate the approach by experiments with a scalable scenario of Bridge play.

Verification of Strategic Ability

In this section we provide an overview of the relevant variants of **ATL**. We refer the to (Alur, Henzinger, and Kupferman 2002; van der Hoek and Wooldridge 2002; Schobbens 2004; Bulling and Jamroga 2011; Jamroga 2015) for details.

Models, Strategies, Outcomes

A *concurrent epistemic game structure* or *CEGS* is given by $M = \langle \mathbb{A}\mathrm{gt}, St, Props, V, Act, d, o, \{\sim_a | a \in \mathbb{A}\mathrm{gt}\}\rangle$ which includes a nonempty finite set of all agents $\mathbb{A}\mathrm{gt} = \{1, \ldots, k\}$, a nonempty set of states St, a set of atomic propositions $Props$ and their valuation $V : Props \to 2^{St}$, and a nonempty finite set of (atomic) actions Act. Function $d : \mathbb{A}\mathrm{gt} \times St \to 2^{Act}$ defines nonempty sets of actions available to agents at each state, and o is a (deterministic) transition function that assigns the outcome state $q' = o(q, \alpha_1, \ldots, \alpha_k)$ to state q and a tuple of actions $\langle \alpha_1, \ldots, \alpha_k\rangle$ for $\alpha_i \in d(i, q)$ and $1 \leq i \leq k$, that can be executed by $\mathbb{A}\mathrm{gt}$ in q. We write $d_a(q)$ instead of $d(a, q)$, and

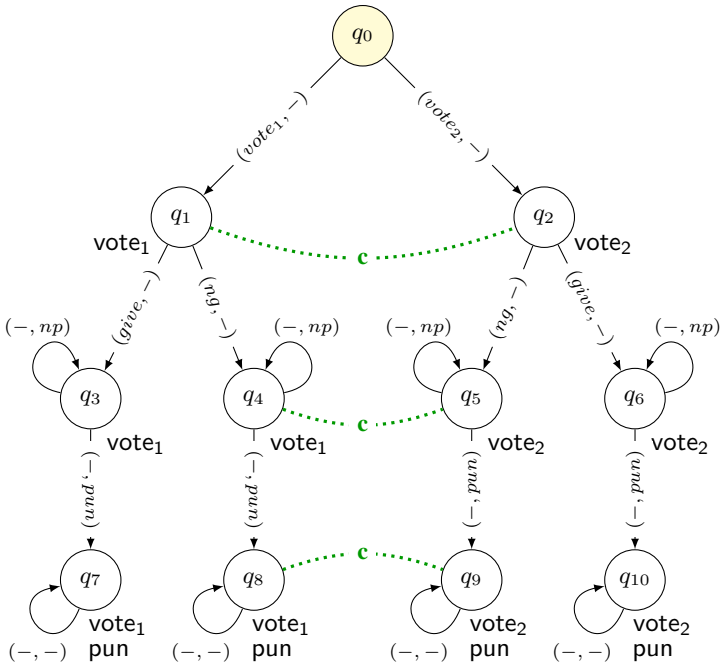

Figure 1: A simple model of voting and coercion

define $d_A(q) = \prod_{a \in A} d_a(q)$ for each $A \subseteq \mathbb{A}\mathrm{gt}, q \in St$. Every $\sim_a \subseteq St \times St$ is an epistemic equivalence relation satisfying $d_a(q) = d_a(q')$ for $q \sim_a q'$.

Example 1. *Consider a very simple voting scenario with two agents: the voter v and the coercer c. The voter casts a vote for a selected candidate $i \in \{1, \ldots, n\}$ (action $vote_i$). For simplicity, we assume that there are only $n = 2$ candidates. Upon exit from the polling station, the voter can hand in a proof of how she voted to the coercer (action $give$) or refuse to hand in the proof (action ng). The proof may be a certified receipt from the election authorities, a picture of the ballot taken with a smartphone, etc. After that, the coercer can either punish the voter (pun) or not punish (np).*

The CEGS M_{vote} modeling the scenario is shown in Figure 1. Proposition vote_i *labels states where the voter has already voted for candidate i. Proposition* pun *indicates states where the voter has been punished. The indistinguishability relation for the coercer is depicted by dotted lines.*

A *strategy* of agent $a \in \mathbb{A}\mathrm{gt}$ is a conditional plan that specifies what a is going to do in each situation. Formally, a *perfect information memoryless strategy* for a can be represented by a function $s_a : St \to Act$ satisfying $s_a(q) \in d(a, q)$ for each $q \in St$. An *imperfect information memoryless strategy* additionally satisfies that $s_a(q) = s_a(q')$ whenever $q \sim_a q'$. Following (Schobbens 2004), we refer to the former as Ir-*strategies*, and to the latter as ir-*strategies*.

A *collective x-strategy* s_A, for coalition $A \subseteq \mathbb{A}\mathrm{gt}$ and strategy type $x \in \{\mathrm{Ir}, \mathrm{ir}\}$, is a tuple of individual x-strategies, one per agent from A. The set of all such strategies is denoted by Σ^x_A. By $s_A|_a$ we denote the strategy of agent $a \in A$ selected from s_A. A partial function $s'_A \colon St \rightharpoonup Act$ is called *a partial x-strategy for A* if $s'_A \subseteq s_A$ for some strategy $s_A \in \Sigma^x_A$. If s'_A and s''_A are partial strategies such that $s'_A \subseteq s''_A$, then we say that s''_A *extends* s'_A.

A *path* $\lambda = q_0 q_1 q_2 \ldots$ is an infinite sequence of states such that there is a transition between each q_i, q_{i+1}. We use $\lambda[i]$ to denote the ith position on path λ (starting from $i = 0$). Function $out(q, s_A)$ returns the set of all paths that can result from the execution of strategy s_A from state q. We will sometimes write $out^{\mathrm{Ir}}(q, s_A)$ instead of $out(q, s_A)$. Moreover, function $out^{\mathrm{ir}}(q, s_A) = \bigcup_{a \in A} \bigcup_{q \sim_a q'} out(q', s_A)$ collects all the outcome paths that start from states that are indistinguishable from q to at least one agent in A.

Alternating-Time Temporal Logic

We use a variant of **ATL** that explicitly distinguishes between perfect and imperfect information abilities. Formally, the syntax is defined by the following grammar:

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

where $x \in \{\mathrm{Ir}, \mathrm{ir}\}$, $p \in Props$ and $A \subseteq \mathbb{A}\mathrm{gt}$. We read $\langle\!\langle A \rangle\!\rangle_{\mathrm{ir}} \gamma$ as *"A can identify and execute the right strategy to enforce γ,"* $\mathbf{X}$ as *"in the next state,"* $\mathbf{G}$ as *"now and always in the future,"* and $\mathbf{U}$ as *"until."* $\langle\!\langle A \rangle\!\rangle_{\mathrm{Ir}} \gamma$ can be read as *"A might be able to bring about γ if allowed to make lucky guesses along the way."* We focus on the kind of ability expressed by $\langle\!\langle A \rangle\!\rangle_{\mathrm{ir}}$. The other strategic modality (i.e., $\langle\!\langle A \rangle\!\rangle_{\mathrm{Ir}}$) will prove useful when approximating $\langle\!\langle A \rangle\!\rangle_{\mathrm{ir}}$.

The semantics of **ATL** can be defined as follows:

- $M, q \models p$ iff $q \in V(p)$,
- $M, q \models \neg\varphi$ iff $M, q \not\models \varphi$,
- $M, q \models \varphi \wedge \psi$ iff $M, q \models \varphi$ and $M, q \models \psi$,
- $M, q \models \langle\!\langle A \rangle\!\rangle_x \mathbf{X}\varphi$ iff there exists $s_A \in \Sigma^x_A$ such that for all $\lambda \in out^x(q, s_A)$ we have $M, \lambda[1] \models \varphi$,
- $M, q \models \langle\!\langle A \rangle\!\rangle_x \mathbf{G}\varphi$ iff there exists $s_A \in \Sigma^x_A$ such that for all $\lambda \in out^x(q, s_A)$ and $i \in \mathbb{N}$ we have $M, \lambda[i] \models \varphi$,
- $M, q \models \langle\!\langle A \rangle\!\rangle_x \psi \mathbf{U} \varphi$ iff there exists $s_A \in \Sigma^x_A$ such that for all $\lambda \in out^x(q, s_A)$ there is $i \in \mathbb{N}$ for which $M, \lambda[i] \models \varphi$ and $M, \lambda[j] \models \psi$ for all $0 \leq j < i$.

We will often write $\langle A \rangle \varphi$ instead of $\langle\!\langle A \rangle\!\rangle_{\mathrm{ir}} \mathbf{X}\varphi$ to express one-step abilities under imperfect information. Additionally, we define *"now or sometime in the future"* as $\mathbf{F}\varphi \equiv \top \mathbf{U} \varphi$.

Example 2. *Consider model M_{vote} from Example 1. The following formula expresses that the coercer can ensure that the voter will eventually either have voted for candidate i (presumably chosen by the coercer for the voter to vote for) or be punished:* $\langle\!\langle c \rangle\!\rangle_{\mathrm{ir}} \mathbf{F}(\neg \mathsf{pun} \rightarrow \mathsf{vote}_i)$. *We note that it holds in M_{vote}, q_0 for any $i = 1, 2$. The strategy for c that validates the property is $s_c(q_3) = np$, $s_c(q_4) = s_c(q_5) = s_c(q_6) = pun$ for $i = 1$, and symmetrically for $i = 2$.*

We refer to the syntactic fragment containing only $\langle\!\langle A \rangle\!\rangle_{\mathrm{ir}}$ modalities as $\mathbf{ATL}_{\mathrm{ir}}$, and to the one containing only $\langle\!\langle A \rangle\!\rangle_{\mathrm{Ir}}$ modalities as $\mathbf{ATL}_{\mathrm{Ir}}$.

Proposition 1 ((Alur, Henzinger, and Kupferman 2002; Schobbens 2004; Jamroga and Dix 2006))**.** *Model checking $\mathbf{ATL}_{\mathrm{Ir}}$ is $\mathbf{P}$-complete and can be done in time $O(|M| \cdot |\varphi|)$ where $|M|$ is the number of transitions in the model and $|\varphi|$ is the length of the formula.*

Model checking $\mathbf{ATL}_{\mathrm{ir}}$ is $\mathbf{\Delta_2^P}$-complete wrt $|M|$ and $|\varphi|$.

Reasoning about Knowledge

Having indistinguishability relations in the models, we can interpret knowledge modalities K_a in the standard way:

- $M, q \models K_a\varphi$ iff $M, q' \models \varphi$ for all q such that $q \sim_a q'$.

The semantics of *"everybody knows" (E_A)* and *common knowledge (C_A)* is defined analogously by assuming the relation $\sim_A^E = \bigcup_{a\in A} \sim_a$ to aggregate individual uncertainty within A, and that $\sim_A^C$ is the transitive closure of $\sim_A^E$. Additionally, we assume $\sim_\emptyset^E$ to be the minimal reflexive relation. We will also use $[q]_\mathcal{R} = \{q' \mid q\mathcal{R}q'\}$ to denote the image of q wrt relation $\mathcal{R}$.

Example 3. *The following formulae hold in M_{vote}, q_0 for any $i = 1, 2$ by virtue of strategy s_c presented in Example 2:*

- $\langle\!\langle c\rangle\!\rangle_{\text{ir}} \mathbf{F}\big((\neg K_c \mathsf{vote_i}) \to \mathsf{pun}\big)$*: The coercer has a strategy so that, eventually, the voter is punished unless the coercer has learnt that the voter voted as instructed;*
- $\langle\!\langle c\rangle\!\rangle_{\text{ir}} \mathbf{G}\big((K_c \mathsf{vote_i}) \to \neg\mathsf{pun}\big)$*: Moreover, the coercer can guarantee that if he learns that the voter obeyed, then the voter will not be punished.*

Alternating Epistemic Mu-Calculus

It is well known that the modalities in $\mathbf{ATL}_{\text{Ir}}$ have simple fixpoint characterizations (Alur, Henzinger, and Kupferman 2002), and hence $\mathbf{ATL}_{\text{Ir}}$ can be embedded in a variant of μ-calculus with $\langle\!\langle A\rangle\!\rangle_{\text{Ir}}\mathbf{X}$ as the basic modality. At the same time, the analogous variant of μ-calculus for imperfect information has incomparable expressive power to $\mathbf{ATL}_{\text{ir}}$ (Bulling and Jamroga 2011), which means that, under imperfect information, $\mathbf{ATL}$ and fixpoint specifications provide an inherently different view of ability.

Formally, the syntax of *alternating epistemic μ-calculus ($\mathbf{AE}\mu\mathbf{C}$)* takes the next-time fragment of $\mathbf{ATL}_{\text{ir}}$, possibly with epistemic modalities, and adds the least fixpoint operator μ. The greatest fixpoint operator ν is defined as dual. We only consider the alternation-free fragment of $\mathbf{AE}\mu\mathbf{C}$ ($\mathbf{af}$-$\mathbf{AE}\mu\mathbf{C}$) here. The semantics is standard, and we omit it due to lack of space.

Example 4. *Consider the $\mathbf{AE}\mu\mathbf{C}$ formula $\mu Z.\big((\neg\mathsf{pun} \to \mathsf{vote_i}) \vee \langle c\rangle Z\big)$, i.e., the "naive" fixpoint translation of the formula $\langle\!\langle c\rangle\!\rangle_{\text{ir}}\mathbf{F}\big(\neg\mathsf{pun} \to \mathsf{vote_i}\big)$ from Example 2. The fixpoint computation produces the whole set of states St. Thus, in particular, $M_{vote}, q_0 \models \mu Z.\big((\neg\mathsf{pun} \to \mathsf{vote_i}) \vee \langle c\rangle Z\big)$.*

Proposition 2 ((Bulling and Jamroga 2011))**.** *Model checking $\mathbf{af}$-$\mathbf{AE}\mu\mathbf{C}$ with strategic modalities for up to 2 agents is $\mathbf{P}$-complete and can be done in time $O(|\sim|\cdot|\varphi|)$ where $|\sim|$ is the size of the largest equivalence class among $\sim_1, \ldots, \sim_k$, and $|\varphi|$ is the length of the formula.*

For coalitions of size at least 3, the problem is between $\mathbf{NP}$ and $\mathbf{\Delta_2^P}$ wrt $|\sim|$ and $|\varphi|$.

Lower Bounds for Strategic Ability

The complexity of $\mathbf{AE}\mu\mathbf{C}$ model checking seems much more attractive than that of $\mathbf{ATL}_{\text{ir}}$. Unfortunately, the expressivity results imply that there is no fixpoint translation that captures *exactly* the meaning of $\mathbf{ATL}_{\text{ir}}$ operators. It might be possible, however, to come up with a translation tr that provides a *lower bound* of the actual strategic abilities, i.e., such that $M, q \models tr(\langle\!\langle A\rangle\!\rangle_{\text{ir}}\gamma)$ implies $M, q \models \langle\!\langle A\rangle\!\rangle_{\text{ir}}\gamma$. In other words, a translation which can only reduce, but never enhance the abilities of the coalition.

We begin by investigating the "naive" fixpoint translation, and show that it works in some cases, but not in general. Then, we propose how to alter the semantics of the next-time modality so that a general lower bound can be obtained. We focus first on reachability goals, expressed by formulae $\langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi$, and then move on to the other modalities.

Trying It Simple for Reachability Goals

We start with the simplest translation, analogous to that of (Alur, Henzinger, and Kupferman 2002): $tr_1(\langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi) = \mu Z.(\varphi \vee \langle A\rangle Z)$. Unfortunately, this translation provides neither a lower nor an upper bound, cf. the counterexamples in (Bulling and Jamroga 2014).

Proposition 3. *$M, q \models \mu Z.(\varphi \vee \langle A\rangle Z)$ does not universally imply $M, q \models \langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi$. The converse implication does not hold either.*

Consider now a slightly stronger fixpoint specification: $tr_2(\langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi) = \mu Z.(E_A\varphi \vee \langle A\rangle Z)$. This new translation works to an extent, as the following proposition shows.

Proposition 4.

1. *$M, q \models \mu Z.(E_\emptyset\varphi \vee \langle\emptyset\rangle Z)$ iff $M, q \models \langle\!\langle\emptyset\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi$;*
2. *If $|A| = 1$, then $M, q \models \mu Z.(E_A\varphi \vee \langle A\rangle Z)$ implies $M, q \models \langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi$, but the converse does not hold;*
3. *If $|A| > 1$, then $M, q \models \mu Z.(E_A\varphi \vee \langle A\rangle Z)$ does not imply $M, q \models \langle\!\langle A\rangle\!\rangle_{\text{ir}}\mathbf{F}\varphi$, and vice versa.*

Proof. **Case 1:** follows from the fact that for the empty coalition the ir–reachability is equivalent to the IR–reachability, which in turn has a fixpoint characterization.

Case 2: Let us assume that $A = \{a\}$ for some $a \in \mathbb{A}\text{gt}$. We define the sequence $\{F_j\}_{j\in\mathbb{N}}$ of $\mathbf{af}$-$\mathbf{AE}\mu\mathbf{C}$ formulae s.t. $F_0 = K_a\varphi$ and $F_{j+1} = F_0 \vee \langle a\rangle F_j$, for all $j \geq 0$. From Kleene fixed-point theorem we have $[\![\mu Z.(K_a\varphi \vee \langle a\rangle Z)]\!] = \bigcup_{j=0}^\infty [\![F_j]\!]$, where $\{[\![F_j]\!]\}_{j\in\mathbb{N}}$ is a non-decreasing monotone sequence of subsets of St. Now, we prove that for each $j \in \mathbb{N}$ there exists a partial strategy s_a^j s.t. $dom(s_a^j) = [\![F_j]\!]$, $\forall q \in dom(s_a^j)\ \forall\lambda \in out^{\text{ir}}(q, s_a^j)\ \exists k \leq j\ \lambda_k \models \varphi$, and $s_a^j \subseteq s_a^{j+1}$. The proof is by induction on j. We constructively build s_a^{j+1} from s_a^j for each $j \in \mathbb{N}$. The base case is trivial. For the inductive step, firstly observe that for each $j \in \mathbb{N}$ if $q \in [\![F_j]\!]$, then $[q]_{\sim_a} \subseteq [\![F_j]\!]$. As $\sim_a$ is an equivalence relation, for each $q \in [\![F_{j+1}]\!]$ either $[q]_{\sim_a} \subseteq [\![F_j]\!]$ or $[q]_{\sim_a} \subseteq [\![F_{j+1}]\!] \setminus [\![F_j]\!]$. In the first case we put $s_a^{j+1}(q) = s_a^j(q)$. In the second case, we know that there exists a strategy s_a^q s.t. $\forall\lambda \in out^{\text{ir}}(q, s_a^q)\ \lambda_1 \in [\![F_j]\!]$. Moreover, the set of such strategies is shared by the whole class $[q]_{\sim_a}$. We thus put $s_a^{j+1}(q') = s_a^q(q')$ for all $q' \in [q]_{\sim_a}$. We finally define the partial strategy $s_a = \bigcup_{j\in\mathbb{N}} s_a^j$. For each $q \in St$ s.t. $M, q \models \mu Z.(K_a\varphi \vee \langle a\rangle Z)$, either $M, q \models \varphi$, or φ is reached along each path consistent with any extension of s_a

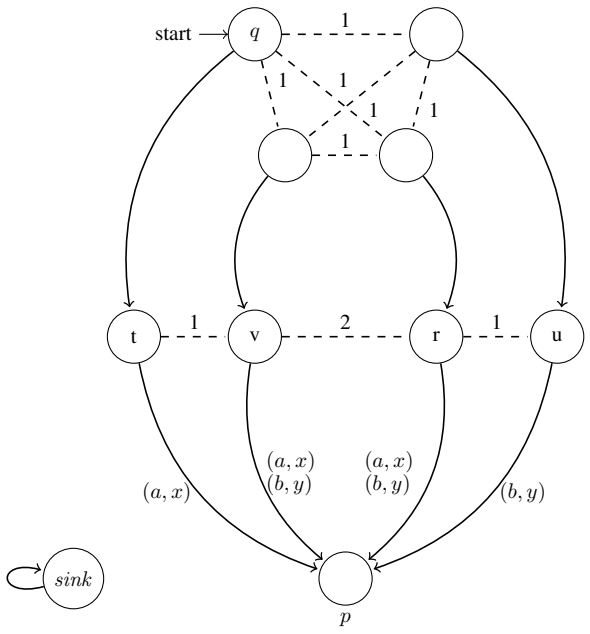

Figure 2: CEGS M_1: a counterexample for reachability

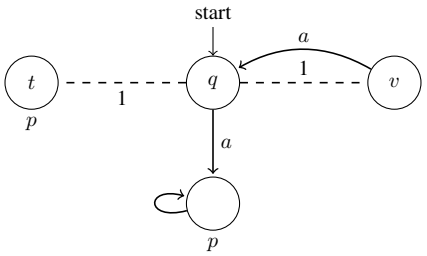

Figure 3: CEGS M_2: the next-step operator limitations

to a full strategy. The invalidity of the converse implication is shown in (Bulling and Jamroga 2011, Proposition 4).

Case 3: Consider the CEGS M_1 presented in Figure 2. We assume that $d_1(s) = \{a, b\}$ and $d_2(s) = \{x, y\}$, for $s \in \{t, v, r, u\}$. In the remaining states the protocols allow only one transition. For clarity, we omit from the figure the transitions leaving the states t, v, r, and u, leading to $sink$. Note now that $M, q \models \mu Z.(E_{\{1,2\}}\varphi \vee \langle\{1,2\}\rangle Z)$ and $M, q \not\models \langle\langle 1, 2\rangle\rangle_{\text{ir}}\mathbf{F}\varphi$. For larger coalitions A, we extend the model with a sufficient number of spurious (idle) agents.

For the other direction, we use the counterexample from (Bulling and Jamroga 2011, Proposition 4), extended with appropriately many spurious agents. □

As Propositions 3 and 4 show, a natural fixpoint translation provides lower bounds for $\mathbf{ATL}_{\text{ir}}$ verification only in a limited number of instances. Also, the bound is rather loose, as the following example demonstrates.

Example 5. *Consider the single-agent CEGS presented in Figure 3. The sole available strategy, in which agent* 1 *selects always action* a*, enforces eventually reaching* p*, i.e.,* $M_2, q \models \langle\langle 1\rangle\rangle_{\text{ir}}\mathbf{F}p$*. On the other hand,* $M_2, q \not\models \mu Z.(E_A p \vee \langle 1\rangle Z)$*. This is because the next-step operator in* $\mathbf{ATL}_{\text{ir}}$ *requires reaching* p *in the succeeding state from* all *the states indistinguishable from* q*, whereas* p *is reached from* t, q, v *in zero, one, and two steps, respectively.*

To obtain a tighter lower bound, and one that works universally, we will introduce modality $\langle A\rangle^\bullet$ which can be seen as a semantic variant of the next-step ability operator $\langle A\rangle$ where agents "steadfastly" pursue their goal in a variable number of steps within the indistinguishability class.

Steadfast Next Step Operator

We begin by defining the auxiliary function $Reach$ so that $q \in Reach_M(Q, s_A, \varphi)$ iff $q \in Q$ and all the paths executing s_A from q eventually reach φ without leaving Q, except possibly for the last step:

$$Reach_M(s_A, Q, \varphi) = \{q \in St \mid q \in Q \text{ and } \forall \lambda \in out(q, s_A) \; \exists i \; \forall 0 \leq j < i \; . \; M, \lambda[i] \models \varphi \text{ and } \lambda[j] \in Q\}.$$

The *steadfast next-step operator* $\langle A\rangle^\bullet$ is defined as follows:

- $M, q \models \langle A\rangle^\bullet\varphi$ iff there exists $s_A \in \Sigma_A^{\text{ir}}$ such that $Reach_M(s_A, [q]_{\sim_A^C}, \varphi) = [q]_{\sim_A^C}$.

Now we can propose our ultimate attempt at the lower bound for reachability goals: $tr_3(\langle\langle A\rangle\rangle_{\text{ir}}\mathbf{F}\varphi) = \mu Z.(E_A\varphi \vee \langle A\rangle^\bullet Z)$, with the following result.

Proposition 5. *If* $M, q \models \mu Z.(E_A\varphi \vee \langle A\rangle^\bullet Z)$*, then* $M, q \models \langle\langle A\rangle\rangle_{\text{ir}}\mathbf{F}\varphi$*. The converse does not universally hold.*

Proof. The proof is similar to the proof of Proposition 4. As previously, let us define the sequence $\{F_j\}_{j\in\mathbb{N}}$ of $\mathbf{af}$-$\mathbf{AE}\mu\mathbf{C}$ formulae s.t. $F_0 = E_A\varphi$ and $F_{j+1} = F_0 \vee \langle A\rangle^\bullet F_j$, for all $j \geq 0$. We use the derived sequence $\{H_j\}_{j\in\mathbb{N}}$ s.t. $H_j = \langle A\rangle^\bullet F_j$ for all $j \in \mathbb{N}$. From Kleene fixed-point theorem we have $[\![\mu Z.(E_A\varphi \vee \langle A\rangle^\bullet Z)]\!] = [\![F_0]\!] \cup \bigcup_{j=0}^{\infty}[\![H_j]\!]$. Observe that, as $\sim_C^A$ is an equivalence relation, for each $q \in St$ and $j \in \mathbb{N}$, if $[q]_{\sim_C^A} \cap [\![H_j]\!] \neq \emptyset$, then $[q]_{\sim_C^A} \subseteq [\![H_j]\!]$.

We prove that for each $j \in \mathbb{N}$ there exists a partial strategy s_A^j s.t. $dom(s_A^j) = [\![H_j]\!]$, $\forall q \in dom(s_A^j)$ $\forall \lambda \in out^{\text{ir}}(q, s_A^j)$ $\exists k \in \mathbb{N}$ $\lambda_k \models E_A\varphi$, and $s_A^j \subseteq s_A^{j+1}$. The proof is by induction on j. In the base case of $H_0 = \langle A\rangle^\bullet E_A\varphi$ observe that if $q \in [\![H_0]\!]$ then there exists a partial strategy $s_A^{0,q}$ with $dom(s_A^{0,q}) = [q]_{\sim_C^A}$ s.t. every $\lambda \in out^{\text{ir}}(q, s_A^{0,q})$ stays in $[q]_{\sim_C^A}$ until it reaches a state where $E_A\varphi$ holds. We can now define $s_A^0 = \bigcup_{[q]_{\sim_C^A} \in St/\sim_C^A} s_A^{0,q}$, where any choice of the representative from a given abstraction class is correct. For the inductive step, we divide the construction of s_A^{j+1} in two cases. Firstly, if $q \in [\![H_j]\!]$, then we put $s_A^{j+1}(q) = s_A^j(q)$. Secondly, let $q \in [\![H_{j+1}]\!] \setminus [\![H_j]\!]$. In this case there exists a partial strategy $s_A^{j+1,q}$ with $dom(s_A^{j+1,q}) = [q]_{\sim_C^A}$ s.t. each outcome $\lambda \in out^{\text{ir}}(q, s_A^{j+1,q})$ stays in $[q]_{\sim_C^A}$ until it reaches a state q' s.t. either $q' \models E_A\varphi$ or $q' \in [\![H_j]\!]$. In the latter, from the inductive assumption we know that following s_A^{j+1} always leads to reaching $E_A\varphi$ without leaving $[\![H_j]\!]$. We thus

$s_A^{j+1} = \bigcup_{[q]_{\sim_C^A} \in St/\sim_C^A} s_A^{j+1,q}$, where, again, any choice of the representative from an abstraction class is correct.

Finally, we can build a partial strategy $s_A = \bigcup_{j\in\mathbb{N}} s_A^j$, whose any extension is s.t. for each $q \in St$, if $M, q \models \mu Z.(E_A\varphi \vee \langle A\rangle^\bullet Z)$, then a state in which $E_A\varphi$ holds is eventually reached along each outcome $\lambda \in out^{\mathrm{ir}}(q, s'_A)$. This concludes the proof. □

Thus, tr_3 indeed provides a universal lower bound for reachability goals expressed in $\mathbf{ATL}_{\mathrm{ir}}$. Moreover, for $|A| = 1$, the bound is strictly tighter than the one provided by tr_2:

Proposition 6. *For $A = \{a\}$, if $M, q \models \mu Z.(K_a\varphi \vee \langle a\rangle Z)$, then $M, q \models \mu Z.(K_a\varphi \vee \langle a\rangle^\bullet Z)$. The converse does not universally hold.*

Proof. It suffices to observe that $q \models \langle a\rangle\varphi$ implies $\langle a\rangle^\bullet\varphi$, for any $\varphi \in \mathbf{af}\text{-}\mathbf{AE}\mu\mathbf{C}$. Note that this is true only for single-agent coalitions. For the converse, notice that in CEGS M_2 from Figure 3 we have $q \not\models \mu Z.(K_a\varphi \vee \langle a\rangle^\bullet Z)$ and $q \not\models \mu Z.(K_a\varphi \vee \langle a\rangle Z)$. □

Lower Bounds for "Always" and "Until"

So far, we have concentrated on reachability goals. We now extend the main result to all the modalities of $\mathbf{ATL}_{\mathrm{ir}}$:

Theorem 7.

1. *If $M, q \models \nu Z.(C_A\varphi \wedge \langle A\rangle^\bullet Z)$, then $M, q \models \langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \mathbf{G}\varphi$;*
2. *If $M, q \models \mu Z.\big(E_A\varphi \vee (C_A\psi \wedge \langle A\rangle^\bullet Z)\big)$, then $M, q \models \langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \psi\, \mathbf{U}\, \varphi$.*

Proof. We start with the first case. Let us define the sequence $\{G_j\}_{j\in\mathbb{N}}$ of formulae s.t. $G_0 = C_A\varphi$ and $G_{j+1} = G_0 \wedge \langle A\rangle^\bullet G_j$, for all $j \geq 0$. From Kleene fixed-point theorem we have $[\![\langle\!\langle A\rangle\!\rangle \mathbf{G}\varphi]\!] = \bigcap_{j=0}^{\infty} [\![G_j]\!]$. It suffices to prove that for each $j \in \mathbb{N}$ there exists a strategy s_A^j s.t. $\forall q \in [\![G_j]\!]\ \forall \lambda \in out^{\mathrm{ir}}(q, s_A^j)\ \forall 0 \leq k \leq j\ \lambda_k \models \varphi$. This proof is by induction on j, with the trivial base case. Assume that the inductive assumption holds for some $j \in \mathbb{N}$. From the definition of the steadfast next-step operator we can define for each equivalence class $[q]_{\sim_C^A} \in [\![G_{j+1}]\!]/\sim_C^A$ a partial strategy $s_A^{[q]_{\sim_C^A}, j+1}$ s.t. $\forall q \in [q]_{\sim_C^A}\ \forall \lambda \in out^{\mathrm{ir}}(q, [q]_{\sim_C^A})\ \lambda_1 \in [\![G_j]\!]$. We now build $s_A^{j+1} = \bigcup_{[q]_{\sim_C^A} \in [\![G_{j+1}]\!]/\sim_C^A} s_A^{[q]_{\sim_C^A}, j+1} \cup s^j_{A|[\![C_A\varphi]\!]\setminus[\![G_j]\!]}$. Intuitively, s_A^j enforces that a path leaving each $q \in [\![G_{j+1}]\!]$ stays within $[\![C_A\varphi]\!]$ for either infinite number of steps (it then visits $[\![G_j]\!]$ infinitely often) or at least j number of steps in $[\![G_j]\!] \setminus [\![G_{j+1}]\!]$. Note that the correctness of the above definition stems from the fact that $\sim_C^A$ is an equivalence relation.

The proof of case (2) is analogous to Proposition 5. □

Approximation Semantics for $\mathbf{ATL}_{\mathrm{ir}}$

Observe that $M, q \models \langle\!\langle A\rangle\!\rangle_{\mathrm{ir}}\gamma$ always implies $M, q \models E_A\langle\!\langle A\rangle\!\rangle_{\mathrm{Ir}}\gamma$. Based on this, and the lower bounds established in Theorem 7, we propose the *lower approximation* tr and the *upper approximation* TR for $\mathbf{ATL}_{\mathrm{ir}}$ as follows:

$$
\begin{array}{l}
tr(p) = p, \quad tr(\neg\varphi) = \neg TR(\varphi), \quad tr(\varphi \wedge \psi) = tr(\varphi) \wedge tr(\psi),\\
tr(\langle A\rangle\varphi) = \langle A\rangle tr(\varphi),\\
tr(\langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \mathbf{G}\varphi) = \nu Z.(C_A tr(\varphi) \wedge \langle A\rangle^\bullet Z),\\
tr(\langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \psi\, \mathbf{U}\, \varphi) = \mu Z.\big(E_A tr(\varphi) \vee (C_A tr(\psi) \wedge \langle A\rangle^\bullet Z)\big).
\end{array}
$$

$$
\begin{array}{l}
TR(p) = p, \qquad TR(\neg\varphi) = \neg tr(\varphi),\\
TR(\varphi \wedge \psi) = TR(\varphi) \wedge TR(\psi),\\
TR(\langle A\rangle\varphi) = E_A\langle\!\langle A\rangle\!\rangle_{\mathrm{Ir}} \mathbf{X}\, TR(\varphi),\\
TR(\langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \mathbf{G}\varphi) = E_A\langle\!\langle A\rangle\!\rangle_{\mathrm{Ir}} \mathbf{G}\, TR(\varphi),\\
TR(\langle\!\langle A\rangle\!\rangle_{\mathrm{ir}} \psi\, \mathbf{U}\, \varphi) = E_A\langle\!\langle A\rangle\!\rangle_{\mathrm{Ir}}\, TR(\psi)\, \mathbf{U}\, TR(\varphi).
\end{array}
$$

Theorem 8. *For any $\mathbf{ATL}_{\mathrm{ir}}$ formula φ:*

$$M, q \models tr(\varphi) \;\Rightarrow\; M, q \models \varphi \;\Rightarrow\; M, q \models TR(\varphi).$$

Proof. A routine induction on the structure of φ. □

The following is relatively straightforward, and we omit the proof due to lack of space:

Proposition 9. *If φ includes only coalitions of size at most 1, then model checking $tr(\varphi)$ and $TR(\varphi)$ can be done in time $O(|M| \cdot |\varphi|)$. In the general case, the problem is between $\mathbf{NP}$ and $\mathbf{\Delta_2^P}$ wrt $\max_{A\in\varphi}(|\sim_A^C|)$ and $|\varphi|$.*

Thus, our approximations potentially offer computational advantage when we consider coalitions whose members have similar knowledge, and especially when verifying abilities of individual agents.

Experimental Evaluation

Theorem 8 and Proposition 9 validate the approximation semantics theoretically. In this section, we back up the theoretical results by looking at how well the approximations work in practice. We address two issues: the *performance* and the *accuracy* of the approximations.

Benchmarks from literature. Typical classes of models used to estimate the performance of $\mathbf{ATL}_{\mathrm{ir}}$ model checking are **TianJi** (Raimondi 2006; Busard et al. 2014) and **Castles** (Pilecki, Bednarczyk, and Jamroga 2014). Unfortunately, the players in both classes are highly forgetful; in particular, they do not store in their local states some observations that are essential for planning of their remaining play. This means that $\mathbf{AE}\mu\mathbf{C}$ approximations will not be very useful, because formulae of $\mathbf{AE}\mu\mathbf{C}$ specify abilities that are *recomputable* in the sense that the players can, at any stage, compute the remaining part of the winning strategy from the available observations (Bulling and Jamroga 2011). More importantly, we usually do *not* want to assume agents to forget relevant information. This is especially true in security applications, cf. our simple voting example. When verifying existence of coercion strategies, we should model the attacker as shrewd and cunning. In the remainder, we propose and employ a novel benchmark that shares much structural characteristics with the voting scenario.

New benchmark: Bridge card play. We use Bridge play scenarios of a type often considered in Bridge handbooks and magazines. The task is to find a winning strategy for the declarer, usually depicted at the South position (**S**), in the k-endplay of the game, see Figure 4 for an example. The deck consists of $4n$ cards in total (n in each suit),[1] and the initial

[1] In real Bridge, $n = 13$.

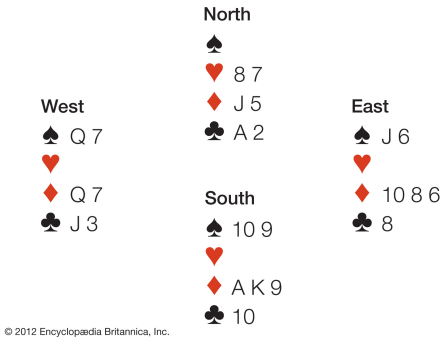

Figure 4: Example 6-endplay in Bridge:

(n,k)	#states	tgen	Lower approx.		Upper approx.		Match	Exact
			tverif	%true	tverif	%true		tg+tv
$(1,1)$	11	0.0005	0.0001	100%	7e-05	100%	100%	0.14
$(2,2)$	310	0.017	0.002	60%	0.001	60%	100%	2.42 h*
$(3,3)$	12626	0.92	0.16	70%	0.05	70%	100%	timeout
$(4,4)$	534722	41.66	172.07	60%	2.61	60%	100%	timeout
$(5,5)^\star$	2443467	2641.86	76 h	100%	1929	100%	100%	timeout

Figure 5: Experimental results: solving endplay in Bridge

(n,k)	#states	tgen	Lower approx.		Upper approx.		Match	Exact
			tverif	%true	tverif	%true		tg+tv
$(1,1)$	19	0.001	0.0003	100%	0.0003	100%	100%	14.93 h*
$(2,2)$	774	0.07	0.01	40%	0.02	50.00%	90%	timeout
$(3,3)$	51865	6.71	29.31	65%	2.45	85%	80%	timeout

Figure 6: Experimental results for absentminded declarer

state captures each player holding k cards in their hand, after having played $n-k$ cards. This way we obtain a family of models, parameterized by the possible values of (n,k). A NoTrump contract is being played; the declarer wins if she takes more than $k/2$ tricks in the endplay. The declarer handles her own cards and the ones of the dummy (**N**). The opponents (**W** and **E**) handle their own hands each. The cards of the dummy are visible to everybody; the other hands are only seen by their owners. Each player remembers the cards that have already been played, including the ones that were used up before the initial state of the k-endplay.

The results of experiments for formula $\varphi \equiv \langle\langle \mathbf{S} \rangle\rangle_{\mathrm{ir}} \mathbf{F}\mathsf{win}$ are shown in Figure 5. The columns present the following information: parameters of the model (n,k), size of the state space (#states), generation time for models (tgen), time and output of verification (tver, %true) for model checking the lower approximation $tr(\varphi)$, and similarly for the upper approximation $TR(\varphi)$; the percentage of cases where the bounds have matched (match), and the total running time of the exact $\mathbf{ATL}_{\mathrm{ir}}$ model checking for φ (tg+tv). The times are given in seconds, except where indicated.

The computation of the lower and upper approximations was done with a straightforward implementation (in Python 3) of the fixpoint model checking algorithm for $\mathbf{AE}\mu\mathbf{C}$ and $\mathbf{ATL}_{\mathrm{ir}}$, respectively. We used the explicit representation of models, and the algorithms were not optimized in any way. The exact $\mathbf{ATL}_{\mathrm{ir}}$ model checking was done with MCMAS 1.2.2 in such a way that the underlying CEGS of the ISPL code was isomorphic to the explicit models used to compute approximations. We ran the experiments with MCMAS for up to 48h per instance. All the tests were conducted on a PC with an Intel Core i5-2500 CPU with dynamic clock speed of 3.30 GHz up to 3,60 GHz, 8 GB of RAM (two modules DDR3, 1600 MHz bus clock), and Windows 10 (64bit). The results in each row are averaged over 20 randomly generated instances, except for ($\star$) where only 1 instance was used.

Discussion of results. In the experiments, our approximations offered a dramatic speedup. Exact model checking of φ was infeasible except for the simplest models (hundreds of states), even with an optimized symbolic model checker like MCMAS. In contrast, the bounds were verified for models up to millions of states. Moreover, our approximations obtained an astonishing level of accuracy: the bounds matched in 100% of the analyzed instances, thus producing fully conclusive output. We suspect that this was partly because we only considered endplays in relatively small decks. We expect the gap to grow for decks of more than 20 cards.

Bridge endplay with absentminded declarer. In the Bridge endplay models, the players always see when a move is made. Thus, for singleton coalitions, the steadfast next-time operator $\langle a \rangle^\bullet$ coincides with the standard next-time abilities expressed by $\langle a \rangle$. In order to better assess the performance of our lower bound, we have considered a variant of the scenario where the declarer is absentminded and does not see the cards being laid on the table until the end of each trick. Moreover, she can play her and the dummy's cards at any moment, even in parallel with the opponents. This results in larger indistinguishability classes for **S**, but also in a general increase of the numer of states and transitions.

The results of the experiments are shown in Figure 6. Note that, for this class of models, the bounds do not match as tightly as before. Still, the approximation was conclusive in an overwhelming majority of instances. Moreover, it grossly outperformed the exact model checking which was (barely) possible only in the trivial case of $n = 1$.

Conclusions

Verification of strategic properties in scenarios with imperfect information is difficult, both theoretically and in practice. In this paper, we suggest that model checking of logics like $\mathbf{ATL}_{\mathrm{ir}}$ can be in some cases obtained by computing an under- and overapproximation of the $\mathbf{ATL}_{\mathrm{ir}}$ specification, and comparing if the bounds match. In a way, our proposal is similar to the idea of may/must abstraction (Godefroid and Jagadeesan 2002; Ball and Kupferman 2006; Lomuscio and Michaliszyn 2016), only our approximations are obtained by transforming formulae rather than models.

We propose such approximations, prove their correctness, and show that, for singleton coalitions, their values can be computed in polynomial time. We also propose a novel benchmark for experimental validation, based on the card game of Bridge, that shares much characteristics with security scenarios. Finally, we report very promising experimental results, in both performance and accuracy of the output.

References

Ågotnes, T.; Goranko, V.; Jamroga, W.; and Wooldridge, M. 2015. Knowledge and ability. In van Ditmarsch, H.; Halpern, J.; van der Hoek, W.; and Kooi, B., eds., *Handbook of Epistemic Logic*. College Publications. 543–589.

Ågotnes, T. 2004. A note on syntactic characterization of incomplete information in ATEL. In *Procedings of Workshop on Knowledge and Games*, 34–42.

Alur, R.; Henzinger, T. A.; and Kupferman, O. 2002. Alternating-time Temporal Logic. *Journal of the ACM* 49:672–713.

Ball, T., and Kupferman, O. 2006. An abstraction-refinement framework for multi-agent systems. In *Proceedings of LICS*, 379–388.

Bulling, N., and Jamroga, W. 2011. Alternating epistemic mu-calculus. In *Proceedings of IJCAI-11*, 109–114.

Bulling, N., and Jamroga, W. 2014. Comparing variants of strategic ability: How uncertainty and memory influence general properties of games. *Journal of Autonomous Agents and Multi-Agent Systems* 28(3):474–518.

Busard, S.; Pecheur, C.; Qu, H.; and Raimondi, F. 2014. Improving the model checking of strategies under partial observability and fairness constraints. In *Formal Methods and Software Engineering*, volume 8829 of *Lecture Notes in Computer Science*. Springer. 27–42.

Busard, S.; Pecheur, C.; Qu, H.; and Raimondi, F. 2015. Reasoning about memoryless strategies under partial observability and unconditional fairness constraints. *Information and Computation* 242:128–156.

Dima, C., and Tiplea, F. 2011. Model-checking ATL under imperfect information and perfect recall semantics is undecidable. *CoRR* abs/1102.4225.

Dima, C.; Enea, C.; and Guelev, D. 2010. Model-checking an alternating-time temporal logic with knowledge, imperfect information, perfect recall and communicating coalitions. In *Proceedings of GANDALF*, 103–117.

Godefroid, P., and Jagadeesan, R. 2002. Automatic abstraction using generalized model checking. In *Proceedings of CAV*, volume 2404 of *Lecture Notes in Computer Science*, 137–150. Springer.

Guelev, D., and Dima, C. 2012. Epistemic ATL with perfect recall, past and strategy contexts. In *Proceedings of CLIMA-XIII*, volume 7486 of *Lecture Notes in Computer Science*, 77–93. Springer.

Guelev, D.; Dima, C.; and Enea, C. 2011. An alternating-time temporal logic with knowledge, perfect recall and past: axiomatisation and model-checking. *Journal of Applied Non-Classical Logics* 21(1):93–131.

Huang, X., and van der Meyden, R. 2014. Symbolic model checking epistemic strategy logic. In *Proceedings of AAAI*, 1426–1432.

Jamroga, W., and Dix, J. 2006. Model checking ATL_{ir} is indeed Δ_2^P-complete. In *Proceedings of EUMAS'06*, volume 223 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jamroga, W., and van der Hoek, W. 2004. Agents that know how to play. *Fundamenta Informaticae* 63(2–3):185–219.

Jamroga, W. 2015. *Logical Methods for Specification and Verification of Multi-Agent Systems*. ICS PAS Publishing House.

Lomuscio, A., and Michaliszyn, J. 2016. Verification of multi-agent systems via predicate abstraction against ATLK specifications. In *Proceedings of AAMAS*, 662–670.

Lomuscio, A., and Raimondi, F. 2006. Model checking knowledge, strategies, and games in multi-agent systems. In *Proceedings of AAMAS*, 161–168.

Lomuscio, A.; Qu, H.; and Raimondi, F. 2015. MCMAS: An open-source model checker for the verification of multi-agent systems. *International Journal on Software Tools for Technology Transfer*. Availabe online.

Pilecki, J.; Bednarczyk, M.; and Jamroga, W. 2014. Synthesis and verification of uniform strategies for multi-agent systems. In *Proceedings of CLIMA XV*, volume 8624 of *Lecture Notes in Computer Science*, 166–182. Springer.

Raimondi, F. 2006. *Model Checking Multi-Agent Systems*. Ph.D. Dissertation, University College London.

Schobbens, P. Y. 2004. Alternating-time logic with imperfect recall. *Electronic Notes in Theoretical Computer Science* 85(2):82–93.

van der Hoek, W., and Wooldridge, M. 2002. Tractable multiagent planning for epistemic goals. In Castelfranchi, C., and Johnson, W., eds., *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-02)*, 1167–1174. ACM Press, New York.

van der Hoek, W.; Lomuscio, A.; and Wooldridge, M. 2006. On the complexity of practical ATL model checking. In *Proceedings of AAMAS'06*, 201–208. ACM.