Reasoning about Bias in MASs Verification

AAMAS 2026

Vadim Malvone **Chunyan Mu**

Télécom Paris University of Aberdeen

27th May, 2026





Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

Reasoning about bias/fairness

Motivations

- Bias in *decision-making* systems is a growing concern
  - intelligent systems increasingly support or automate decisions in e.g., hiring, reviewing, justice, healthcare, ...
- Bias in MASs is *subtle* and *dynamic*
  - bias may arise not only from individual agents but from strategic behaviours over time.
- Existing approaches are limited, struggle to capture:
  - strategic behaviour,
  - temporal evolution, and
  - interaction-driven bias.

Reasoning about bias/fairness

Motivations

- Bias in *decision-making* systems is a growing concern
  - intelligent systems increasingly support or automate decisions in e.g., hiring, reviewing, justice, healthcare, ...
- Bias in MASs is *subtle* and *dynamic*
  - bias may arise not only from individual agents but from strategic behaviours over time.
- Existing approaches are limited, struggle to capture:
  - strategic behaviour,
  - temporal evolution, and
  - interaction-driven bias.

Aim

Develop formal framework to reason about bias in dynamic MASs.

Reasoning about Bias: Key ideas and Contributions

Fairness as non-interference

We introduce a novel definition of bias inspired by non-interference, intuitively:

- A system is unbiased if bias-sensitive attributes (e.g., gender, race) do not affect behaviour or outcomes when non-bias attributes are fixed.
- Captured via state and path equivalence.

Why this matters?

- Goes beyond simple outcome comparison.
- Captures implicit, structural, and interaction-driven bias.
- Well-suited to strategic MASs.

Key ideas and Contributions

Bias-aware temporal logic

- We propose bATLand bATL*, extensions of ATL/ATL*, with bias-sensitive operators.
- Enable formal specification of fairness and bias properties under strategies.

Verification guarantees

- We develop model-checking algorithms for both logics
- Complexity results:
  - bATL: PTIME-complete
  - bATL*: 2EXPTIME-complete

Contribution

A unified logical and algorithmic framework for formal bias reasoning in strategic, temporal MASs.

Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

Preliminaries

Concurrent Game Structure (CGS)

CGS $M = \langle \mathrm{Ag}, \mathrm{Ap}, S, s_I, \{\mathrm{Act}_i\}_{i \in \mathrm{Ag}}, d, \delta, V \rangle$ such that:

- $\mathrm{Ag} = \{1, \ldots, m\}$ is a nonempty finite set of agents.
- Ap is a nonempty finite set of atomic propositions (atoms).
- $S \neq \emptyset$ is a finite set of *states*, with a set of *initial states* $s_I \subseteq S$.
- For $i \in \mathrm{Ag}$, Act_i is a finite set of *actions*, $\mathrm{Act} = \bigcup_{i \in \mathrm{Ag}} \mathrm{Act}_i$, and $\mathrm{ACT} = \prod_{i \in Ag} \mathrm{Act}_i$ the set of all joint actions.
- The *protocol function* $d : \mathrm{Ag} \times S \to (2^{\mathrm{Act}} \setminus \{\emptyset\})$
- The *transition function* $\delta : S \times \mathrm{ACT} \to S$ assigns a successor state $s' = \delta(s, \vec{a})$ to each $s \in S$, for every joint action $\vec{a} \in \mathrm{ACT}$ such that $a_i \in d(i, s)$ for every $i \in \mathrm{Ag}$.
- $V : S \to 2^{\mathrm{Ap}}$ is the *labelling function*.

Preliminaries

Syntax: ATL/ATL*

Formulas in ATL*:

$$\begin{aligned} \varphi &::= q \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\langle \Gamma \rangle\rangle \psi \\ \psi &::= \varphi \mid \neg\psi \mid \psi \wedge \psi \mid \bigcirc\psi \mid (\psi \mathbf{U} \psi) \end{aligned}$$

Formulas in the ATL fragment are obtained from defn of ATL*:

$$\psi ::= \bigcirc\varphi \mid (\varphi \mathbf{U} \varphi) \mid (\varphi \mathbf{R} \varphi)$$

Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

Modelling Bias

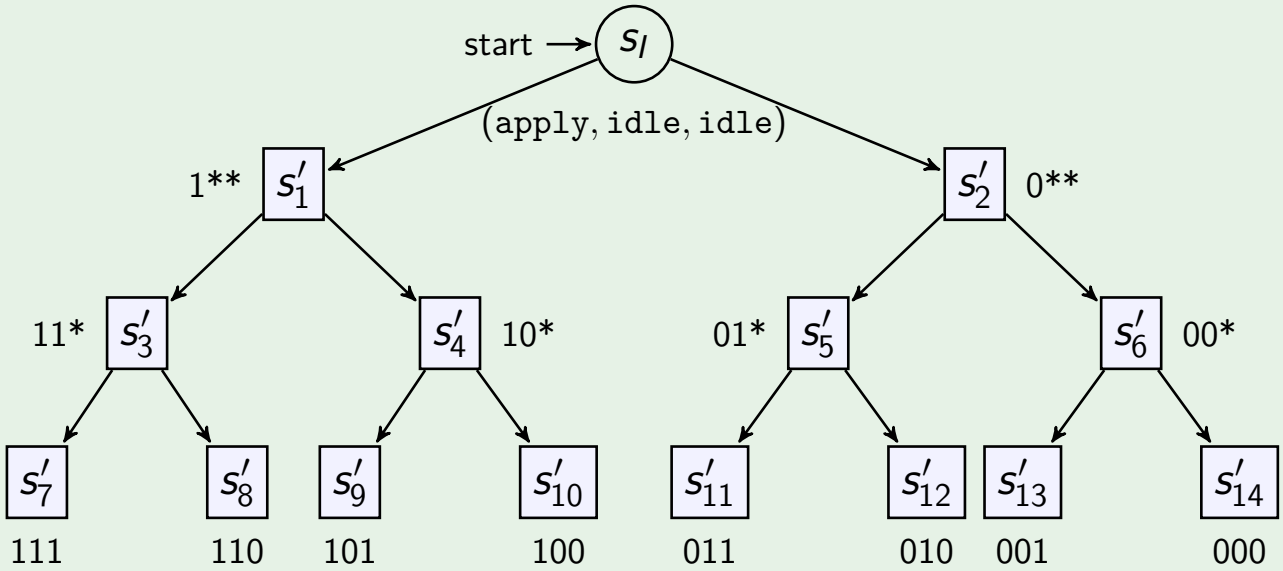
Bias-sensitive Model

A bias-sensitive model is a tuple of the form $\mathcal{G}=(M,\eta)$, where:

- M is a concurrent game structure (CGS);
- $\eta : \mathrm{Ap} \mapsto \mathcal{L}$ is the bias labelling function mapping each atomic proposition to a bias level in $\mathcal{L}$ representing the complete lattice of the bias levels.

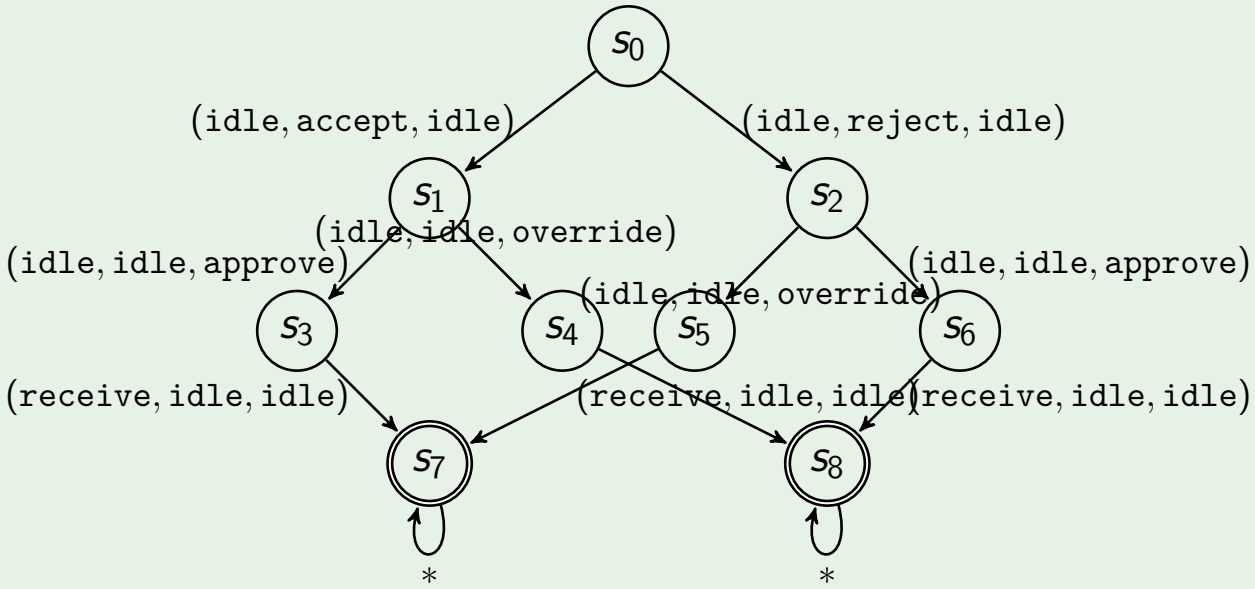
Modelling Bias

Example: college admission system



Modelling Bias

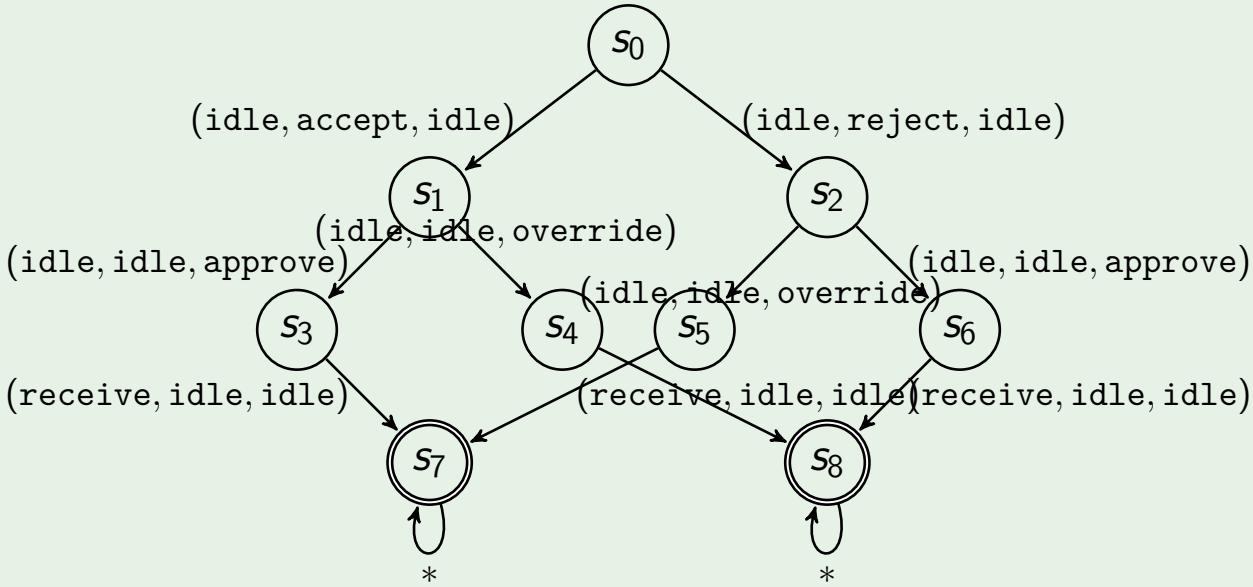
Example: college admission system



$s_0 = \{s_i' \mid 7 \leq i \leq 12\}$: an application received for evaluation.
s_1 and s_2: AI reviewer recommending "accept" and "reject".
s_3 and s_6: human reviewer confirming that of AI reviewer.
s_4 and s_5: human reviewer "overrides" AI reviewer's recommendation.
s_7 and s_8: the final admission outcomes "accept" and "reject".

Modelling Bias

Example: college admission system



Example paths with input $\texttt{merit} = 1$, $\texttt{origin} = 1$, $\texttt{group} = 0$:

$$\pi_1 = (s_8' \in s_0) \to s_2 \to s_6 \to s_8 \; \pi_2 = (s_8' \in s_0) \to s_2 \to s_5 \to s_7$$

and example paths with input $\texttt{merit} = 1$, $\texttt{origin} = 1$, $\texttt{group} = 1$:

$$\pi_3 = (s_7' \in s_0) \to s_1 \to s_3 \to s_7 \; \pi_4 = (s_7' \in s_0) \to s_1 \to s_4 \to s_8$$

Formalising Bias

Bias levels

- We assign bias-sensitive levels with ordering to state atomic propositions, through $\eta : \mathsf{Ap} \mapsto \mathcal{L}$, where $\mathcal{L}$ denotes the complete lattice of the bias-sensitive levels.
- The powerset of atomic state propositions therefore forms a complete lattice $\mathcal{L}$, whose partial ordering is regarding to the bias-sensitive levels of the atomic propositions Ap:

$$\forall v, v' \in \mathsf{Ap},\ v \preceq v' \ \textit{iff}\ \eta(v) \leq \eta(v')$$

Formalising Bias

Non-Bias Equivalent States

We say state $s, s' \in S$ are $\mathcal{NB}$-equivalent, written as $s =_{\mathcal{NB}} s'$, iff:

$$\forall x \in \mathsf{Ap}.\eta(x) \leq \mathcal{NB} \Rightarrow (x \in V(s) \wedge x \in V(s')) \vee (x \notin V(s) \wedge x \notin V(s')).$$

Example: NB-equivalent states

let $\eta(\texttt{merit}) = \eta(\texttt{origin}) = \eta(\texttt{decision}) = \mathcal{NB}$, and $\eta(\texttt{group}) = \mathcal{B}$, and $\mathcal{NB} \preceq \mathcal{B}$, i.e., demographic group is bias-sensitive attribute while merit, origin, and final decision are non-bias-sensitive. We have:

$$s'_7 =_{\mathcal{NB}} s'_8,\ s'_9 =_{\mathcal{NB}} s'_{10},\ s'_{11} =_{\mathcal{NB}} s'_{12},\ s'_{13} =_{\mathcal{NB}} s'_{14}.$$

Formalising Bias

Non-Bias Equivalent Paths

Given a model $\mathcal{G}$, and two paths $\pi, \pi' \in \text{Path}_{\mathcal{G}}$. We say that two paths π and π' are weakly $\mathcal{NB}$-equivalent, written as $\pi \sim_{\mathcal{NB}_w} \pi'$, iff:

- If $|\pi| < |\pi'|$: $\pi_0 =_{\mathcal{NB}} \pi'_0 \Rightarrow \exists j \geq |\pi|.\forall k \geq j.(last(\pi) =_{\mathcal{NB}} \pi'_k)$;
- If $|\pi'| < |\pi|$: $\pi_0 =_{\mathcal{NB}} \pi'_0 \Rightarrow \exists j \geq |\pi'|.\forall k \geq j.(\pi_k =_{\mathcal{NB}} last(\pi'))$;
- Else $|\pi| = |\pi'|$: $\pi_0 =_{\mathcal{NB}} \pi'_0 \Rightarrow last(\pi) =_{\mathcal{NB}} last(\pi')$.

π and π' are strongly $\mathcal{NB}$-equivalent, written as $\pi \sim_{\mathcal{NB}_s} \pi'$, iff:

$$(|\pi| = |\pi'|) \wedge (\pi_0 =_{\mathcal{NB}} \pi'_0 \Rightarrow \forall j \in (0, |\pi|].(\pi_j =_{\mathcal{NB}} \pi'_j)).$$

Formalising Bias

Example: NB-equivalent paths

Consider example paths with input $\texttt{merit} = 1$, $\texttt{origin} = 1$, $\texttt{group} = 0$:

$$\pi_1 = (s_8' \in s_0) \to s_2 \to s_6 \to s_8$$
$$\pi_2 = (s_8' \in s_0) \to s_2 \to s_5 \to s_7$$

and example paths with input $\texttt{merit} = 1$, $\texttt{origin} = 1$, $\texttt{group} = 1$:

$$\pi_3 = (s_7' \in s_0) \to s_1 \to s_3 \to s_7$$
$$\pi_4 = (s_7' \in s_0) \to s_1 \to s_4 \to s_8$$

Cleary: $\pi_1 \not\sim_{\mathcal{NB}_w} \pi_3$, $\pi_1 \sim_{\mathcal{NB}_w} \pi_4$, $\pi_2 \sim_{\mathcal{NB}_w} \pi_3$, $\pi_2 \not\sim_{\mathcal{NB}_w} \pi_4$.

Formalising Bias

Bias Plicy

Let ψ be an outcome formula and $\mathcal{G} = (\mathsf{Ag}, S, s_I, \mathsf{Act}, \rightarrow, \mathsf{Ap}, L, \eta)$ a model.

- The model $\mathcal{G}$ is said to be weakly biased with respect to ψ, written $\mathcal{G} \Vdash_{\psi} \phi_{\mathcal{B}_{\mathsf{w}}}$, iff

$$\mathsf{Paths}_{\mathcal{G}}(s_I, \psi) = \emptyset \quad \vee \quad \exists \pi, \pi' \in \mathsf{Paths}_{\mathcal{G}}(s_I, \psi) \text{ implies } \pi \not\sim_{\mathcal{NB}_{\mathsf{w}}} \pi'.$$

- The model $\mathcal{G}$ is said to be strongly biased with respect to ψ, written $\mathcal{G} \Vdash_{\psi} \phi_{\mathcal{B}_{\mathsf{s}}}$, iff

$$\mathsf{Paths}_{\mathcal{G}}(s_I, \psi) = \emptyset \quad \vee \quad \exists \pi, \pi' \in \mathsf{Paths}_{\mathcal{G}}(s_I, \psi) \text{ implies } \pi \not\sim_{\mathcal{NB}_{\mathsf{s}}} \pi'.$$

Formalising Bias

Example: Bias Plicy

- Let $\psi = \Diamond(\texttt{decision})$, note that:
- there exists $\pi_1, \pi_3 \in \text{Paths}_{\mathcal{G}}(s_I, \psi)$ s.t. $\pi_1 \not\sim_{\mathcal{NB}_w} \pi_3$, so:
  - $\mathcal{G}$ is a weakly biased model regarding ψ.
- This meets our intuition since:
  - the bias-sensitive attribute of demographic group impacts the final decision while non-bias-sensitive attributes remains the same, i.e.,
  - the final result is influenced by bias-sensitive attributes, throughout the path.

Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

The Bias-Sensitive Logic: bATL*

Syntax: bATL*

$$\varphi ::= q \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\langle \Gamma \rangle\rangle \psi \mid \langle\langle \Gamma \rangle\rangle \mathfrak{B}_{\mathsf{w}} [\psi] \mid \langle\langle \Gamma \rangle\rangle \mathfrak{B}_{\mathsf{s}} [\psi]$$
$$\psi ::= \varphi \mid \neg\psi \mid \psi \wedge \psi \mid \bigcirc\psi \mid (\psi \mathbf{U} \psi)$$

Example

- $\neg\langle\langle A \rangle\rangle \mathfrak{B}_{\mathsf{w}}[\Diamond(\texttt{policyPassed})]$ captures that bias-sensitive factors (e.g., media influence, institutional discrimination) prevent party A from ensuring the policy's success through any strategy;
- $\langle\langle \Gamma \rangle\rangle \mathfrak{B}_{\mathsf{w}}[\Box\ \neg\texttt{error}]$ states that under weak bias consideration, coalition Γ can ensure that no execution path leads to an error state.

Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

Model Checking bATL

Verification problem for bATL formulae

Given a bias-sensitive model $\mathcal{G}=(\mathrm{Ag}, \mathrm{Ap}, S, s_I, \{\mathrm{Act}_i\}_{i\in\mathrm{Ag}}\delta, V, \eta)$ and a bATL formula φ, determine whether it is the case that every initial state $s \in s_I$ satisfies φ.

Theorem

The model checking problem for bATL in our model is decidable, and can be done in PTIME-complete.

Theorem

The model checking problem for bATL* in CGSs is decidable, and can be done in 2EXPTIME-complete.

Outline

1. Motivations and overview
2. Preliminaries: CGS and ATL
3. Modelling Bias
4. The Bias-Sensitive Logic
5. Model Checking bATLand bATL*
6. Conclusion and Future Works

Conclusion and Future Works

Summary

- A novel notion of bias and a corresponding bias policy for MASs,
- bATL logic framework upon a game structure model.
- Verification precess for reasoning about bias within this framework.

Future works

- Extend our framework to incorporate quantitative measures of bias.
- Investigate how strategic interactions among agents affect bias.
- Integrating strategic elements into our logical framework to better capture the complexities of decision-making in competitive and cooperative environments.