UNIVERSITY OF TRENTO - Italy

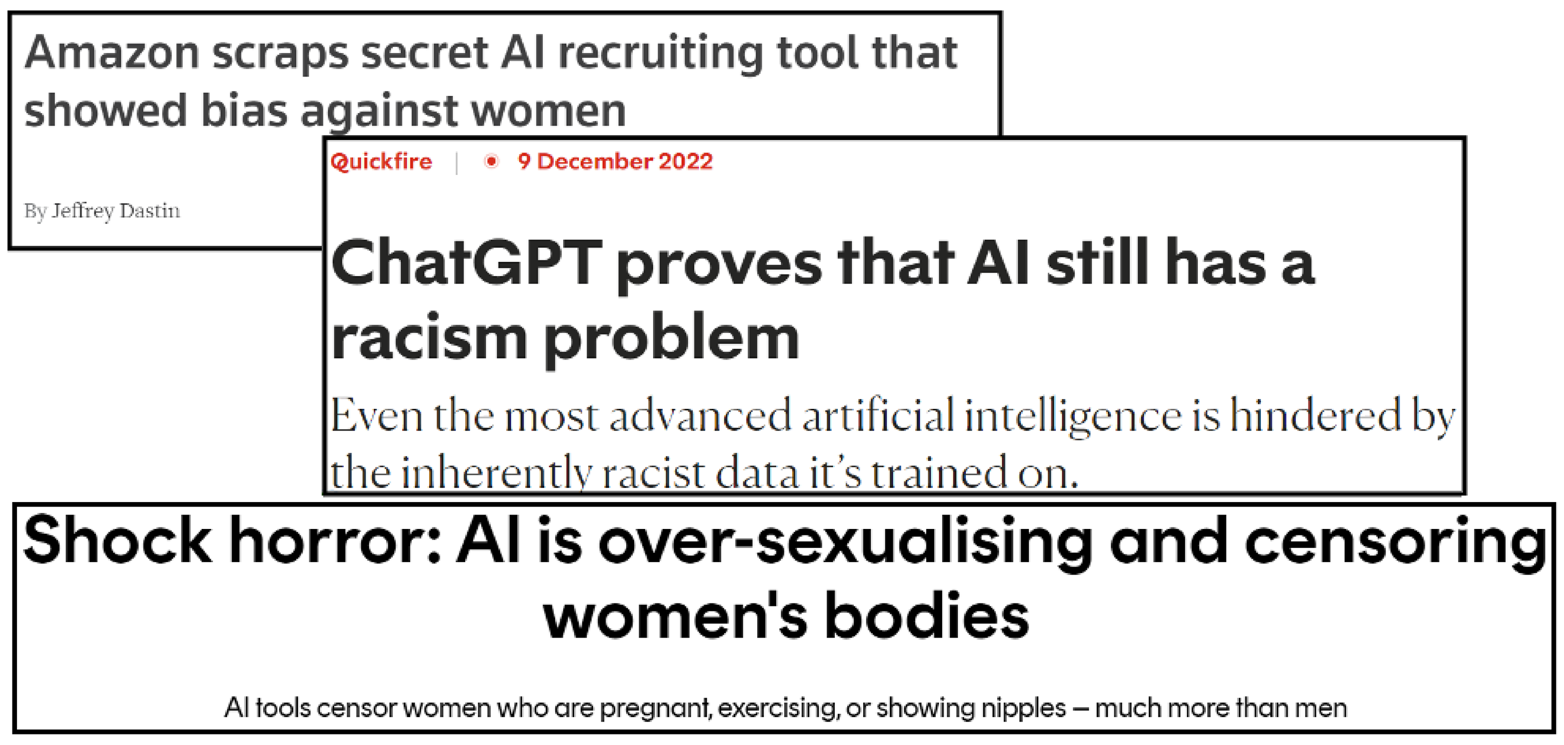
Efficient SMT-based Weighted Model Integration for AI verification

Giuseppe Spallitta[1] Gabriele Masina[1] Paolo Morettin[2] Andrea Passerini[1] Roberto Sebastiani[1]

[1]University of Trento [2]KU Leuven

Code and benchmarks available at: `https://github.com/unitn-sml/wmi-pa`

Why AI verification is becoming relevant


Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

Quickfire | 9 December 2022

ChatGPT proves that AI still has a racism problem

Even the most advanced artificial intelligence is hindered by the inherently racist data it's trained on.

Shock horror: AI is over-sexualising and censoring women's bodies

AI tools censor women who are pregnant, exercising, or showing nipples – much more than men

1. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
2. https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem
3. https://www.cosmopolitan.com/uk/reports/a42915373/ai-discriminates-women/

- Developing neural networks (NN) to help humans in their everyday jobs is becoming easier and easier...
- But are we sure that they learned what we expect?

It is essential to **formally verify** properties on neural network and probabilistic models before releasing them on market:

- Fairness
- Safety
- Robustness to noise

Weighted Model Integration (WMI)

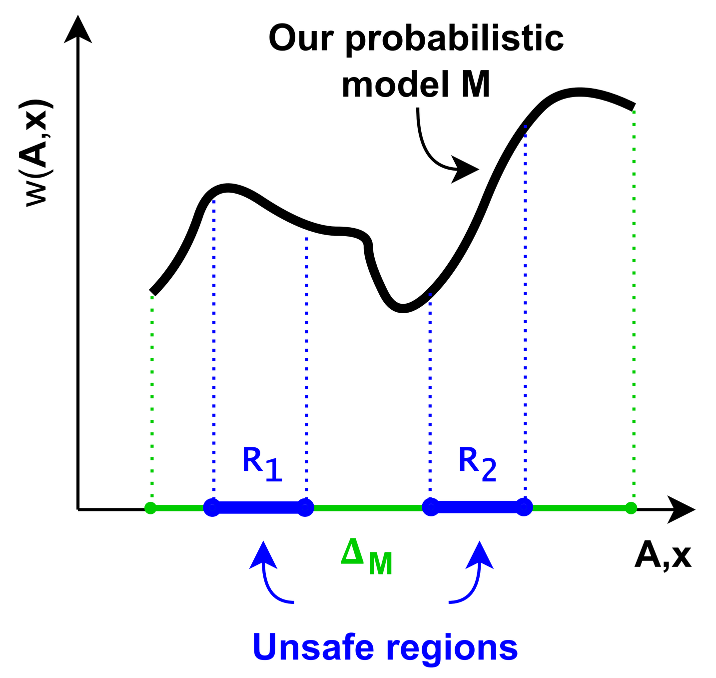
Let $\mathbf{x} \stackrel{\text{def}}{=} \{x_1, ..., x_N\} \in \mathbb{R}^N$ and $\mathbf{A} \stackrel{\text{def}}{=} \{A_1, ..., A_M\} \in \mathbb{B}^M$. $\varphi(\mathbf{x}, \mathbf{A})$ denotes an SMT($\mathcal{LRA}$) formula, while $w(\mathbf{x}, \mathbf{A})$ denotes a non-negative weight function s.t. $\mathbb{R}^N \times \mathbb{B}^M \longmapsto \mathbb{R}^+$.

The **Weighted Model Integral** of $w(\mathbf{x},\mathbf{A})$ over $\varphi(\mathbf{x}, \mathbf{A})$ is:

$$\mathsf{WMI}(\varphi, w|\mathbf{x}, \mathbf{A}) \stackrel{\text{def}}{=} \sum_{\mu^{\mathbf{A}} \in \mathbb{B}^M} \mathsf{WMI_{nb}}(\varphi_{[\mu^{\mathbf{A}}]}, w_{[\mu^{\mathbf{A}}]}|\mathbf{x})$$

$$\mathsf{WMI_{nb}}(\varphi, w|\mathbf{x}) \stackrel{\text{def}}{=} \int_{\varphi(\mathbf{x})} w(\mathbf{x})\, d\mathbf{x}$$

Hybrid probabilistic inference through WMI



$$\Pr(R_1 \vee R_2 \,|\, M) = \frac{\mathsf{WMI}(\Delta_M \wedge (R_1 \vee R_2), w)}{\mathsf{WMI}(\Delta_M, w)}$$

Main issues: can we compute it **efficiently**?

- Integrals are hard to compute...
- The more complex the models, the higher the number of integrals we have to deal with!

Key idea

We propose **SA-WMI-PA**, a novel WMI approach based on partial enumeration which is **aware of the conditional structure of** w.

- Compile the weight function into a **valid formula that drives the enumeration** of assignments in WMIPA.
  $\Longrightarrow$ Takes the best out of KC and WMIPA approaches, the current state-of-the-art frameworks.

Structure-aware WMI-PA

Generate a formula from w, $\mathsf{sk}(w)$, s.t.:

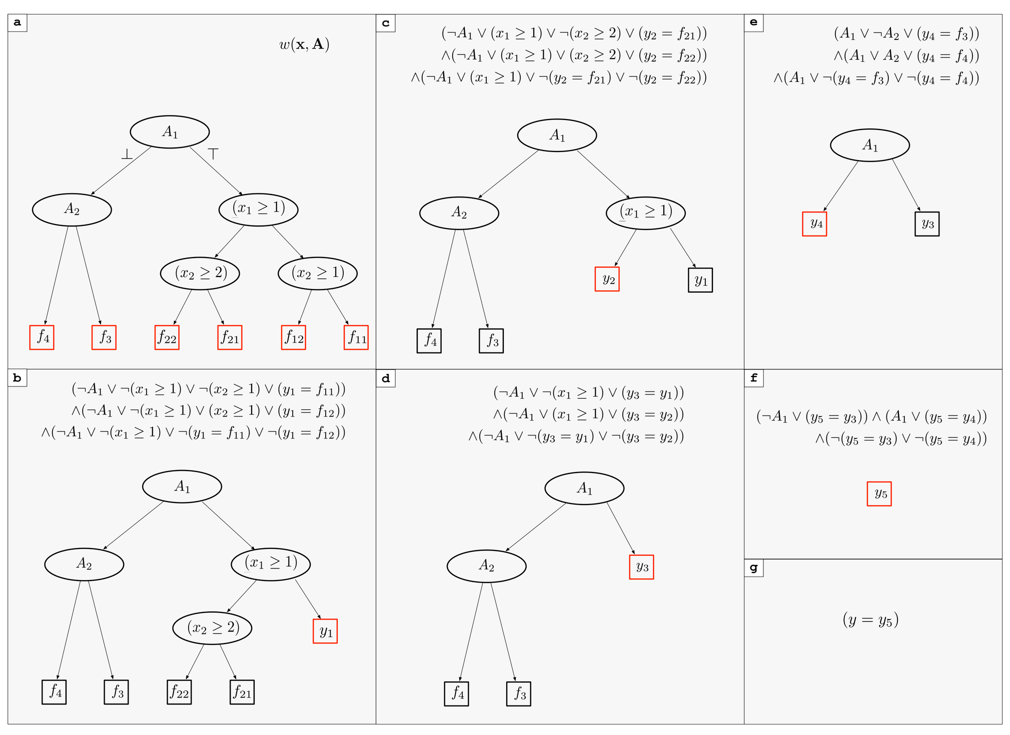
- Its atoms are all and only conditions in w;
- φ is equivalent to $\varphi \wedge \mathsf{sk}(w)$
- Any *partial* truth value assignment μ to the conditions of w ensure $w_{[\mu]}$ is arithmetically computable.

Main advantages:

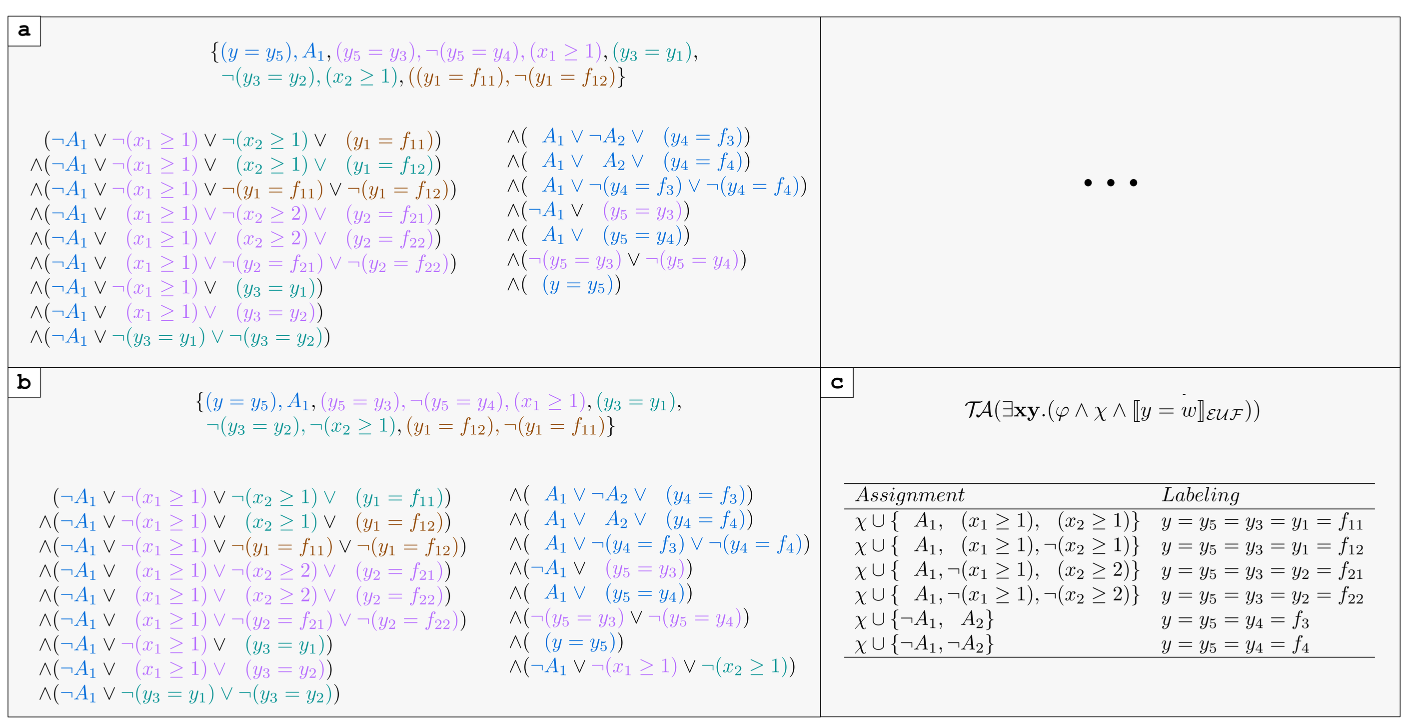
- The SMT solver is not forced to enumerate all conditions on w if they do not impact the final value of $w_{[\mu]}$;
- We can perform partial enumeration on the Boolean atoms.

$\Rightarrow$ **Fewer integrals to compute**

Bottom-up procedure rewriting the **skeleton** of w into a $\mathcal{LRA} \cup \mathcal{EUF}$ formula, $[\![y = w]\!]_{\mathcal{EUF}}$.

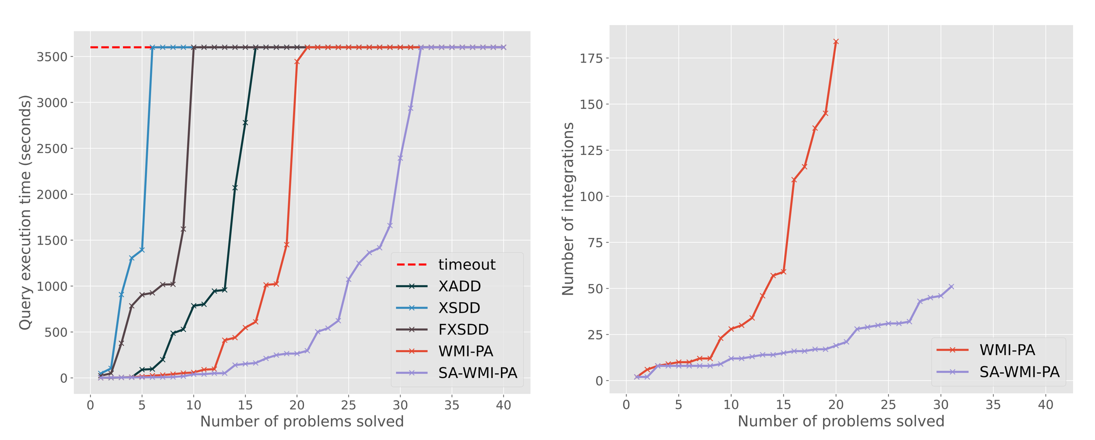


Perform enumeration on the formula $\varphi \wedge [\![y = w]\!]_{\mathcal{EUF}}$



Experimental results

Synthetic problems



Probabilistic inference on DETs