Counting Worlds Branching Time Semantics for Post-Hoc Bias Mitigation in Generative AI

Alessandro G. Buda[1], Giuseppe Primiero[2], Leonardo Ceragioli[2] and Melissa Antonelli[3]

[1]University School for Advanced Studies IUSS Pavia, Piazza della Vittoria n. 15, Pavia, 27100, Italy
[2]University of Milan, Via Festa del Perdono 7, Milano, 20122, Italy
[3]Universität Tübingen & CFvW, Doblerstraße 33, Tübingen, 72074, Germany

Abstract

Generative AI systems are known to amplify biases present in their training data. While several inference-time mitigation strategies have been proposed, they remain largely empirical and lack formal guarantees. We introduce CTLFS, a branching-time logic designed to reason about bias in series of outputs by a generative AI. CTLFS defines modal operators that, based on a counting worlds semantics, allow us to verify whether the current generation process respects intended properties. These include: satisfaction of an intended probability distribution over a protected attribute; prediction of the likelihood of remaining within acceptable bounds as new outputs are generated; and required output elimination to restore fairness. We illustrate the framework on a toy example of biased image generation.

Keywords

Computational Tree Logic, Counting World Semantics, Generative AI, Bias Mitigation

1. Introduction

The presence of bias in Generative AI is a widely recognized phenomenon, stemming primarily from the use of unbalanced training datasets. Furthermore, such original bias reflecting real world disparities is amplified by several factors related to the different components of AI systems, a phenomenon known as bias amplification [1, 2]. In the case of image generation, one of these factors is the poor representation of certain groups of individuals within the training datasets [3]. Recent surveys provide comprehensive taxonomies of bias sources, evaluation methodologies, and bias mitigation strategies in LLMs [4, 5, 6]. These works categorize mitigation approaches into data-level, model-level, and post-processing techniques, highlighting the growing importance of inference-time control, as retraining is costly and difficult to scale to larger models.

Inference-only and post-hoc mitigation methods aim to reduce bias without retraining foundation models. A number of proposals have been offered in this direction: removing biased activations [7]; token-level filtering mechanisms [8]; vector ensemble methods [9, 10]. Simulation-based analyses further evaluate the effectiveness of such inference-time interventions under controlled bias scenarios [11]. Guardrail architectures provide an external safety layer around LLMs modularly combining content classification, policy enforcement, and fallback mechanisms to ensure trustworthy deployment [12]. Broader reviews discuss layered safety frameworks and architectural constraints designed to intercept unsafe or biased outputs before user delivery [13, 14]. Empirical studies also examine bias in LLM-based bias detection systems themselves, revealing disparities between model judgments and human perception [15].

From the point of view of formal analysis and verification, it is useful to check at any point in the output production whether it is possible, given a prompt with an expected distribution of a protected

8th International Workshop on Artificial Intelligence and fOrmal VERification, Logic, Automata, and sYnthesis. OVERLAY 2026 @ FLoC, July 18–19, Lisbon, Portugal

alessandro.buda@iusspavia.it (A. G. Buda); giuseppe.primiero@unimi.it (G. Primiero); leonardo.ceragioli@unimi.it (L. Ceragioli); melissa.antonelli@uni-tuebingen.de (M. Antonelli)
0000-0002-5547-7514 (A. G. Buda); 0000-0003-3264-7100 (G. Primiero); 0000-0001-5250-9720 (L. Ceragioli); 0009-0006-9072-4847 (M. Antonelli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The last two authors contributed equally to this work.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4324/paper1.pdf

attribute, to determine current bias amplification beyond a given acceptable threshold. When it is already known that the output will become unacceptable, continuing using the model may become useless or harmful. This requires predicting the evolution of the output series as early as possible. Following such verification process, one may want to start a post-hoc bias mitigation process which will work on the current output to remove portions of it in order to bring the distribution back within admissible limits. This requires to reason temporally on series of outputs and their possible extensions. The linear temporal logic LTLFS offered in [16] allows to infer frequencies of events on future states of the output. In the present paper, we present CTLFS, an alternative counting world CTL-style strategy to establish whether, at any current state of a series of outputs, the possible final series will lay behind the acceptable threshold for unbiased outputs, and if so how to intervene to mitigate the situation.

2. A toy example



Figure 1: Original training set, which will inform the output's distribution

Consider a dataset D of 20 images with an unfair gender distribution (e.g., 75% M/25% F, Figure 1). Ignoring any possible amplification, we assume that when D is used to train a model M the output series $O = \{o_1, \ldots, o_n\}$ on the prompt *"Give me a picture of an engineer"* repeated n times will reflect the same gender distribution of D. Knowing the current and expected distribution of the output and its cardinality, the objective is to mitigate O at inference stage to obtain an output subset $O' \subset O$ with cardinality as close to n as possible with a desired gender distribution (e.g., $50\%M/50\%F$). Hence, we formulate the following questions:

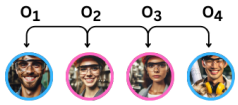


Figure 2: A subset of the output at the fourth item in the series.

1. At output $m < n$, are we still within the intended distribution (e.g. a fair one)? See Figure 2
2. If so, what is the probability to be within the intended distribution at output m' s.t. $m < m' \leq n$?
3 If not, how many outputs do we need to remove to restore it? See Figure 3

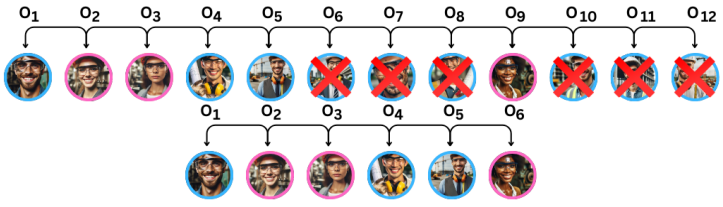


Figure 3: A mitigated output subset.

In the following we will refer to odds for the sake of readability, however it can be easily extended to frequencies and/or probabilities. Consider again Figure 1 as the training set where the initial odds are $15/5 = 3/1$ in favor of males. Given the results up to o_4 (Figure 2), the output is still within the intended (say fair) distribution of the gender property. To establish if the results will still be within

the intended distribution at say o_7, one requires calculating the odds of the output being male at all intermediate steps (given the final expected distribution):

1. At o_5 they are $13/3 > 4/1$;
2. At o_6 they can be $12/3 = 4/1$ if we obtained a male at o_5, or $13/2 > 6/1$ if we obtained a female subject;
3. At o_7 they can be $11/3 < 4/1$ if we obtained a male subject both at o_5 and o_6, or $12/2 = 6/1$ if we obtained a male at o_5, a female at o_6, and *vice versa*, or $13/1$ if we obtained a female subject both at o_5 and o_6.

Consider now Figure 3, where twelve outputs have been generated within the same unfair distribution of the training dataset ($9/3 = 3/1$ in favor of males). At o_5 the result is $3/2$ in favor of male subjects. Given that we have already reached the maximum number of male subjects within the intended distribution for the set of generated outcomes, we should start mitigating by removing all the following outcomes that will result in male engineers.

Our aim is to provide a semantics to reason logically about such distributions, their extensions and required mitigation when the desired output is no longer achieved.

3. The logic CTLFS and its counting semantics

The logic LTLFS is introduced in [16] to reason about series of probabilistic events. We present an extension CTLFS with additional operators and an alternative semantics. In what follows, we will use $p_1, p_2, \ldots$ for atomic propositions. We assume that such atoms are linguistic representations of our probabilistic events; accordingly, as events are mutually exclusive and collectively exhaustive, the occurrence of an event (i.e. an atom in the language) implies the negation of any other event within the same sample space (i.e. negation of an atom is equivalent to the disjunction of all other atoms in the same sample space). For example, for p_1="male", $\neg p_1 \leftrightarrow p_2$="female", or $\neg p_1 \leftrightarrow p_2 \vee p_3$="female or trans". $\phi, \psi, \gamma, \ldots$ denote molecular propositions and $\mathfrak{P}$ denotes the set of propositional variables.

Definition 1 (Language of CTLFS)**.** Formulas of CTLFS are defined by the grammar:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid \phi \vee \phi \mid \Box_q\phi \mid \blacksquare_q\phi \mid \circ_q\phi \mid \triangleright_q\phi$$

$$\psi := \triangledown_q\phi \mid \dagger_q\phi$$

where $p \in \mathfrak{P}$ and $q \in \mathbb{Q}_{[0,1]}$.

In the following, we rely on standard probability theory terminology. Intuitively, an event denotes an output on a prompt; a series of events denotes a series of outputs of size n on a prompt; the outcome or value of an event denotes the occurrence of a given (possibly protected) property in an output. A state validates the occurrence of an event. We now formalize this intuition. ϕ denotes a state formula. A propositional variable p denotes linguistically the value of a protected attribute of output o of model $\mathcal{M}$, and it can be closed under negation (useful for binary protected attributes), conjunction and disjunction. We call this an *event*. Given a series of n events, $\Box_q\phi$ says that at the current state event ϕ has happened *at least* $q \times m$ times, where $m \leq n$ is the number of outputs so far. The formula $\blacksquare_q\phi$ expresses that at the current state event ϕ could have happened (possibly in a different series) at least $q \times m$ times, where $m \leq n$ is the number of outputs so far within a well-defined probability distribution Φ_Σ. The formula $\circ_q\phi$ expresses that, among the possible complete series of outputs, there is one such series at some point of which the event ϕ has happened at least $q \times n$ times, where n is the total number of outputs happening in that series. $\triangleright_q\phi$ expresses that, given a complete series of outputs within a well-defined probability distribution, and a point in that series, the probability that event ϕ happens at the next time in that series is at least q, namely $q \times (n - m)$ times. ψ denotes a path formula. The formula $\triangledown_q\phi$ expresses that, given a complete series of events, event ϕ happens at least $q \times n$ times in that series. The formula $\dagger_q\phi$ expresses that, starting at a given event of a complete series with outputs in a well-defined probability distribution, the probability that such series is completed with outputs still respecting that distribution is at least q.

Definition 2 (Model of CTLFS)**.** Let Σ be a well-defined probability distribution on an event space. We define a model $\mathcal{M} = (W, W^1, W^n, R_b, R_f, \mathcal{V})$, where

- $W = \{w_{1.1}, w_{1.2}, \ldots, w_{1.l}, w_{2.1}, w_{2.2}, \ldots, w_{2.l \times l}, \ldots w_{n.1}, \ldots, w_{n.m}\}$ is a finite set of worlds (or states) s.t. l is the number of possible outcomes for each event, n is the number of events for each series and $m = l^n$ is the number of possible series of events;
- $W^1 \supset W^1 = \{w_{1.1}, w_{1.2}, \ldots, w_{1.l}\}$ is the set of *root worlds*, denoting the set of possible values for the first event of the series, with cardinality l;
- $W^n \supset W^n = \{w_{n.1}, w_{n.2}, \ldots, w_{n.m}\}$ is the set of *leaf worlds*, *i.e.*, the set of possible outcomes for the last event of the series, with cardinality m;
- R_b is an anti-symmetric, beginning, reflexive, and transitive (backward) relation, *i.e.*, a *childhood* relation between different events of the series, such that:
  - $\forall w \in W^1, w' R_b w \iff w' = w$: root events have no parents;
  - for each pair $w_{i.x}, w_{j,y} \in W$ if $w_{i.x} R_b w_{j.y}$ then $i \geq j$: non-root events are children of previous ones
  - for each $w_{i.j} \in W$ with $i > 1$, $\exists! w_{i-1.k} : w_{i.j} R_b w_{i-1.k}$ with $k = \lceil j/l \rceil$, and l is the number of possible outcomes for each event: each non-root event $w_{i.j}$ at time $i > 1$ can have only one *direct* parent world at time $i - 1$, whose secondary index is given by the ceiling function of j/l. Therefore, given a series of n events with l possible outcomes for each event, there exists only one path from each of the possible l^n outcomes for the last event through all events in the series back to the root world.
- R_f is an anti-symmetric, beginning, reflexive, and transitive (forward) relation, *i.e.*, a *parenthood* relation between different events of the series, such that:
  - $\forall w \in W^n, w' R_f w \iff w' = w$: leaf events have no children;
  - for each pair $w_{i.x}, w_{j,y} \in W$, if $w_{i.x} R_f w_{j.y}$ then $i \leq j$: previous events are parents of following ones;
  - for each $w_{i.j} \in W$ with $i < n$, there exists a set $\{w_{i+1.h+1}, w_{i+1.h+2}, \ldots, w_{i+1.h+l}\}$ with $h = [(j-1) \times l]$, and l is the number of possible outcomes for each event, *i.e.*, each non-leaf world $w_{i.j}$ representing one possible outcome for the event at time $i < n$ has l *direct* children at time $i + 1$, with secondary indexes $[(j-1) \times l] + 1, [(j-1) \times l] + 2, \ldots, [(j-1) \times l] + l$. Therefore, given a series of n events, with l possible outcomes for each event, and given a non leaf world $w_{i.j}$, representing the jth possible outcome for the ith event in the series, there are l^{n-i} possible paths from $w_{i.j}$ to the leaf worlds, passing across all the intermediate events.
- $\mathcal{V} : W \mapsto \mathcal{P}(\mathfrak{P})$ is a valuation function assigning to each world in W a subset of $\mathfrak{P}$ holding at that world.

We denote with $W^i = \{w_{i.1}, w_{i.2}, \ldots, w_{i.l}\}$ the set of possible outcomes for the ith event of the series.

Definition 3 (Path)**.** Given W and $R_f \subseteq W \times W$, we define a *path* between events i and j of the series (with $i < j$) as:

$$\pi_{[i.x,j.z]} = \{\langle w_{i.x}, w_{i+1.y}, \ldots, w_{j.z}\rangle : w_{k.v} \in W, (w_{k.v}, w_{k+1.w}) \in R_f, \forall k \in [i, j-1]\}$$

An equivalent definition can be provided using R_b instead of R_f. A *complete path* π^* is a path from the first to the last event of the series, including all the intermediate steps: $\pi^*_{[x,z]} = \pi_{[1.x,n.z]}$.

Definition 4 (Set of paths)**.** Given W and $R_f \subseteq W \times W$, and a world $w_{i.j}$, we define the *set of paths* originating in $w_{i.j}$ as: $\Pi^{[i.j]} = \{\langle w_{i.j}, \ldots, w_{n.k}\rangle, \ldots, \langle w_{i.j}, \ldots, w_{n.m}\rangle\}$. The set $\Pi^* = \{\langle w_{1.1}, \ldots, w_{n.k}\rangle, \ldots, \langle w_{1.l}, \ldots, w_{n.m}\rangle\}$ is the *set of complete paths*.

Definition 5 (Φ_Σ)**.** Given a probability distribution Σ assigning to any $p \in \mathfrak{P}$ the frequency $\mu(p)$, the function $\Phi_\Sigma : \Pi^* \mapsto \mathcal{P}(\Pi^*)$ takes the set of complete paths in W and returns any of its member $\pi^*_{[x,z]}$ such that

$$\frac{|\{w_{i.j} \in \pi^*_{[x,z]} : w_{i.j} \in \mathcal{V}(p)\}|}{|\pi^*_{[x,z]}|} = \mu(p)$$

In other words, every path in the co-domain of Φ_Σ shows for each $p \in \mathfrak{P}$ a set of $\mu(p) \times |\pi^*|$ worlds in which p is true. We denote with $\Pi^{[i.j]}_\Sigma$ the set of paths originating in $w_{i.j}$ returned by Φ_Σ, and with Π^*_Σ the set of complete paths returned by Φ_Σ. We denote with $v : \pi_{[i.x,j.z]} \mapsto \mathcal{P}(\mathfrak{P})$ any function for the series of events corresponding to path $\pi_{[i.x,j.z]}$, identifying one among the possible $m = l^n$ series of events, or possibly complete paths.

Definition 6 (Reachable worlds)**.** Given a model $\mathcal{M}$, for any world $w_{i.j} \in W$ we denote:

- $R_b^{w_{i.j}} = \{w' \in W : w_{i.j} R_b w'\}$, the set of all worlds accessible from $w_{i.j}$ according to the transitive closure on R_b.
- $R_f^{w_{i.j}} = \{w' \in W : w_{i.j} R_f w'\}$, the set of all worlds accessible from $w_{i.j}$ according to the transitive closure on R_f.

Definition 7 (Semantics for state and path formulas)**.** For any CTLFS-formula ϕ, a model $\mathcal{M}$, a world $w \in W$, we define the relation $\mathcal{M}, w \vDash$ inductively as follows:

$$\begin{aligned}
&\mathcal{M}, w_{i.j} \vDash p \iff w_{i.j} \in \mathcal{V}(p)\\
&\mathcal{M}, w_{i.j} \vDash \neg\phi \iff M, w_{i.j} \nvDash \phi\\
&\mathcal{M}, w_{i.j} \vDash \phi_1 \wedge \phi_2 \iff M, w_{i.j} \vDash \phi_1 \text{ and } M, w_{i.j} \vDash \phi_2\\
&\mathcal{M}, w_{i.j} \vDash \phi_1 \vee \phi_2 \iff M, w_{i.j} \vDash \phi_1 \text{ or } M, w_{i.j} \vDash \phi_2\\
&\mathcal{M}, w_{i.j} \vDash \Box_q\phi \iff \frac{|R_b^{w_{i.j}}(\phi)|}{|R_b^{w_{i.j}}|} \geq q.\\
&\mathcal{M}, w_{i.j} \vDash \blacksquare_q\phi \iff \frac{|R_b^{w_{i.k}}(\phi)|}{|R_b^{w_{i.k}}|} \geq q \text{ for some } \pi^*_\Sigma \text{ with} w_{i.k \leqslant \geqslant j} \in \pi^*_\Sigma \text{ and } w_{i.k \leqslant \geqslant j} \in W^i.\\
&\mathcal{M}, w_{i.j} \vDash \circ_q\phi \iff \frac{|R_b^{w_{i.j}}(\phi)|}{|\pi^*|} \geq q \text{ for some } \pi^*.\\
&\mathcal{M}, w_{i.j} \vDash \triangleright_q\phi \iff \frac{|\ R_f^{w_{i+1.k}}(\phi)\ |}{|\ R_f^{w_{i+1.k}}\ |} \geq q \text{ for some } \pi^*_\Sigma \text{ such that } w_{i.j}, w_{i+1,k} \in \pi^*_\Sigma,
\end{aligned}$$

where $R_b^{w_{i.j}}(\phi) = \{w' \in R_b^{w_{i.j}} : \mathcal{M}, w_{i.j} \vDash \phi\}$ and $R_f^{w_{i.j}}(\phi) = \{w' \in R_f^{w_{i.j}} : \mathcal{M}, w_{i.j} \vDash \phi\}$

For any CTLFS-formula ϕ, a model $\mathcal{M}$, a (possibly complete) path $\pi_{[i.x,j.z]}$, we define the relation $\mathcal{M}, \pi_{[i.x,j.z]} \vDash$ inductively as follows:

$$\begin{aligned}
&\mathcal{M}, \pi^*_{[x,z]} \vDash \triangledown_q\phi \iff \frac{|\ R_f^{w_{1.x}}(\phi)\ |}{|\ \pi^*_{[x,z]}\ |} \geq q\\
&\mathcal{M}, \pi_{[i.x,j.z]} \vDash \dagger_q\phi \iff \mathcal{M}, \pi^*_{[x,z]} \supseteq \pi_{[i.x,j.z]} \vDash \triangledown_{q'}\phi \text{ with } \pi^*_{[x,z]} \in \Pi^*_\Sigma \text{ and } \frac{|\ \Pi^{[j.z]}_\Sigma\ |}{|\ \Pi^{[j.z]}\ |} \geq q.
\end{aligned}$$

4. Formalising our example

Example 1. Consider the model in Fig. 4 representing $n = 6$ generations of images on the prompt *"Give me a picture of an engineer"*, with $l = 2$ possible outcomes (M and F, for the sake of simplicity) for each generation, and $m = 2^6 = 64$ possible series of events. Connections between outcomes represent

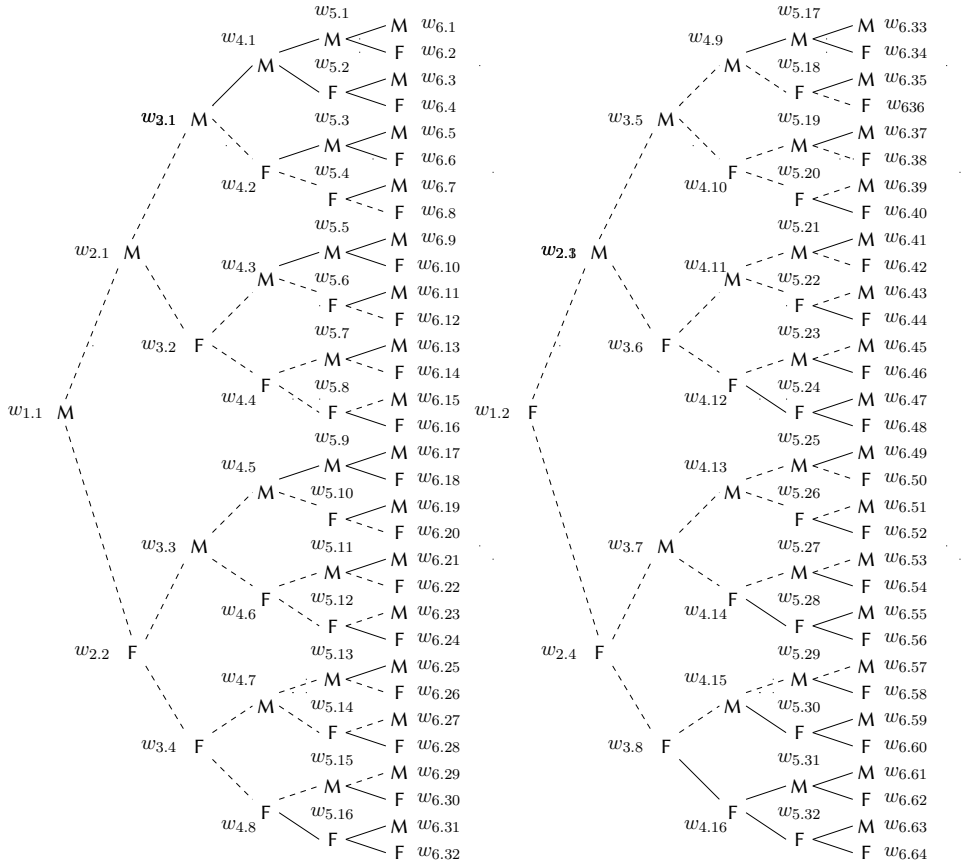


Figure 4: A model representing 6 generations of images on the prompt *"Give me a picture of an engineer"*. Connections represent both the experimental (R_b), and the forward-looking (R_f) relations. Paths in Π^*_Σ are marked with dashed lines.

both the backward- (R_b), and the forward-looking (R_f) relations. Paths with valuations in Φ_Σ, *i.e.*, paths in Π^*_Σ, are marked with dashed lines.

$$W = \{w_{1.1}, w_{1.2}, w_{2.1}, \ldots, w_{2.4}, w_{3.1}, \ldots, w_{3.8}, w_{4.1}, \ldots, w_{4.16}, w_{5.1}, \ldots, w_{5.32}, w_{6.1}, \ldots, w_{6.64}\}$$

Worlds $w_{1.1}$ and $w_{1.2}$ are the root worlds. For each *non-root* world there exists only one direct parent world belonging to the set of possible outcomes for the previous step in the series (*e.g.*, world $w_{3.5}$ has only one direct parent world, $w_{(3-1).\lceil 5/2 \rceil} = w_{2.3}$), and there is only one complete path for each world connecting this to the corresponding root world (*e.g.*, from $w_{3.5}$ to $w_{2.3}$, and then to $w_{1.2}$). The subset of W $\{w_{6.1}, \ldots, w_{6.64}\}$ is the set of leaf worlds. For each non leaf world there are $l = 2$ direct children worlds, representing the possible outcomes for the next event of the series (e.g. $w_{2.3}$ has two direct children, $w_{2+1.[(3-1)\times 2]+1} = w_{3.5}$, and $w_{2+1.[(3-1)\times 2]+2} = w_{3.6}$), and there are $2^{6-2} = 16$ possible paths from $w_{2.3}$ to the corresponding leafs (from $w_{2.3}$ to $w_{6.33}, w_{6.34}, w_{6.35}, \ldots w_{6.48}$).

Example 2. Consider again the model in Figure 4. The following hold:

- $w_{1.1} \vDash M$ because $w_{1.1}$ is in $\mathbf{a}(M)$, *i.e.*, M is true in $w_{1.1}$
- $w_{1.1} \vDash \neg F$ because $w_{1.1} \nvDash F$, *i.e.*, F is false in $w_{1.1}$
- $w_{3.1} \vDash \Box_1 M$ because $|R_b^{w_{3.1}}(M)| \,/\, |R_b^{w_{3.1}}| = 1$
- $w_{3.2} \vDash \Box_{2/3} M$ because $|R_b^{w_{3.1}}(M)| \,/\, |R_b^{w_{3.1}}| = 2/3$
- $w_{3.1} \vDash \blacksquare_1 F$ and $w_{3.2} \vDash \blacksquare_1 F$ because there exists $\pi^*_\Sigma = \{\langle w_{1.2}, \ldots, w_{6.57}\rangle\}$ and there exists $w_{3.8}$ belonging to both π^*_Σ and W^3 such that $|R_b^{w_{3.8}}(F)| \,/\, |R_b^{w_{3.8}}| = 1$
- $w_{3.1} \vDash \circ_{3/6} M \wedge \circ_{0/6} F$ because there exists π^* , *e.g.*, $\{\langle w_{[1.1]} \ldots w_{[6.1]}\rangle\}$, such that $w_{3.1}$ is in π^*, $|R_b^{w_{3.1}}(M)| \,/\, |\pi^*| = 3/6$, and $|R_b^{w_{3.1}}(F)| \,/\, |\pi^*| = 0$

- $w_{3.1} \vDash \triangleright_1 F$ because there exists $\pi^*_\Sigma = \{\langle w_{1.1}, \ldots, w_{6.8}\rangle\}$ such that $w_{3.1}, w_{4.2}$ are in π^*_Σ and $|R_f^{w_{4.2}}(F)| \,/\, |R_f^{w_{4.2}}| = 1$
- $w_{3.2} \vDash \triangleright_{2/3} F$ because there exists $\pi^*_\Sigma = \{\langle w_{1.1}, \ldots, w_{6.12}\rangle\}$ such that $w_{3.2}, w_{4.3}$ are in π^*_Σ and $|R_f^{w_{4.3}}(F)| \,/\, |R_f^{w_{4.3}}| = 2/3$
- $\pi^*_{[2,63]} \vDash \triangledown_{1/6} M$ because starting from the root world $w_{1.2}$ we can reach, by means of R_f across the path $\pi^*_{[2,63]}$, only one world over six in which M holds, formally $|R_f^{w_{1.2}}(M)| \,/\, |\pi^*_{[2,63]}| = 1/6$
- $\pi_{[1.1,3.1]} \vDash \dagger_{1/8} M$ because $\pi_{[1.1,3.1]}$ is one of the paths originating in $w_{1.1}$ returned by Φ_Σ, and there is only one possible path over eight starting from $w_{3.1}$ returned by Φ_Σ.

Example 3. The results depicted in Figure 2 are represented in the model in Figure 4 by the path starting from $w_{1.1}$ (thus we can ignore the set of paths $\Pi^{[1.2]}$), passing across $w_{[2.2]}$ and $w_{[3.4]}$, and reaching $w_{[4.7]}$. This sub-series of event is still within a fair distribution, in fact we are in a path in Π^*_Σ (a dashed path in the figure). Using the defined language: $w_{4.7} \vDash \Box_{1/2} M \wedge \Box_{1/2} F$. Both M and F have appeared in proportion $\geq 1/2$ among the ancestors, hence the output is balanced so far.

At $w_{4.7}$ we can look at the probability of completing the series within a fair distribution. The path $\pi_{[1.1,4.7]}$ is one of the paths originating from $w_{1.1}$ returned by Φ_Σ, and from $w_{4.7}$ there are $l^{n-i} = 2^2 = 4$ possible continuations, of which only 2 yield a Σ-compatible completion reaching $3M/3F$ at event 6. Since $|\Pi^{[4.7]}_\Sigma| \,/\, |\Pi^{[4.7]}| = 2/4 = 1/2$, we have: $\pi_{[1.1,4.7]} \vDash \dagger_{1/2} M$. This means that, being currently on a Σ-compatible path, the probability of completing the series within the fair distribution is exactly $1/2$. Any single deviation at step 5 or 6 eliminates all remaining Σ-compatible completions.

The two Σ-compatible completions can be further characterised using the forward-looking operator $\triangleright$. They are $w_{4.7} \to w_{5.13}(M) \to w_{6.26}(F)$ and $w_{4.7} \to w_{5.14}(F) \to w_{6.27}(M)$. In both cases $|R_f^{w_{5,k}}(\phi)| \,/\, |R_f^{w_{5,k}}| = 1/2$, yielding: $w_{4.7} \vDash \triangleright_{1/2} F \;\wedge\; \triangleright_{1/2} M$. Both next outputs preserve a fair completion, but each forces a unique outcome at step 6: the series is balanced but offers no redundancy.

Consider now the mitigation scenario depicted in Figure 3, where 12 outputs are generated following the training distribution ($9/3 = 3/1$ in favour of males). Since at most 3 female outputs can appear in the full series, the largest achievable balanced subset has cardinality 6, yielding $|\pi^*| = 6$ for the fair-subset path. At o_5 the ratio is $3/2$ in favour of males: as $|R_e^{w_{5,j}}(M)| = 3 = 1/2 \times |\pi^*|$, we have: $w_{5,j} \vDash \circ_{1/2} M$, meaning the male quota for the balanced subset is exhausted. Mitigation is nonetheless possible: there exists a Σ'-compatible path with target distribution $50\%M/50\%F$ passing through $w_{5,j}$ such that $\pi^*_{\Sigma'} \vDash \nabla_{1/2} F$, confirming that a $3M/3F$ subset is achievable from our current position. The mitigation strategy consists in removing all subsequent male outputs (marked $\times$ in Figure 3), yielding a balanced subset $O' \subset O$ of 6 outputs ($\pi_{[1.1,6.26]}$ in Figure 4).

5. Conclusions

Our modalities are inspired by the counting semantics from [17] and its use for a pragmatic of information [18], by probabilistic interpreted systems [19], by probability logics [20, 21, 22], as well as by the counting operators introduced in counting logics, which were mostly developed in connection with complexity [23, 24, 25] and finite model theory [26, 27]. The strategy of checking againts admissible distributions is at the basis of a proof-theoretic and possible-worlds semantics for trustworthiness evaluation and preservation of non-deterministic computations [28, 29, 30]

Several directions for future work are open. First, the current model treats all transitions within a given probability assignement as equally probable. A natural extension would assign explicit weights to the transitions, reflecting prior knowledge about the generation distribution. This would allow a more fine-grained account of the likelihood of individual outputs, and could be integrated into the semantics of the counting operators via a weighted reachability measure, enriching the expressive power of the logic without altering its overall structure. Second, a more significant extension concerns the case in which the reference probability distribution is not fixed in advance, but is refined at runtime as new outputs are observed. Rather than assuming the Markov chain as a known input, one could start from an initial distribution and update the transition probabilities each time a new output is produced: if a "male" output is generated, the weight on the transition towards output "female" is

slightly increased for the subsequent step, progressively steering the generation towards the intended distribution. This setting is closely related to the study of parametric Markov chains [31], where transition probabilities are expressed as functions of a set of parameters that can be synthesised or updated to satisfy a given specification [32], and to recent work on the formal verification of Markov processes with learned parameters [33]. The resulting framework would be an adaptive model in which the counting worlds structure is progressively identified from observations. Establishing formal guarantees on the convergence of such updates and on the correctness of the verification and mitigation procedures in this dynamic setting constitutes an open problem of independent interest. Finally, a software implementation of the checking and mitigation procedures described in this paper is under development for the MIRAI Toolbox, which has already been applied to several classification tasks [34, 35, 36]. The tool will take as input a stream of generative AI outputs, a target distribution and a fairness threshold, and will return either a certificate of compliance or a mitigation plan specifying which outputs to remove. This will allow the framework to be evaluated on real-world systems, including both text-to-image models and large language models.

Acknowledgments

This research was funded by EU—Next Generation EU, Mission 4, Component 1, CUP I13C24000400008, by Ministero dell'Università e della Ricerca (MUR) through PRIN 2022 Project SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin (20223E8Y4X), and Project "Departments of Excellence 2023-2027" awarded to the Department of Philosophy "Piero Martinetti" of the University of Milan which provided funding for a research visit in 2025 to the last author.

Declaration on Generative AI

During the preparation of this work, the authors used DALL-E (OpenAI) for figures 1, 2 and 3 in order to: Generate images. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Seshadri, S. Singh, Y. Elazar, The bias amplification paradox in text-to-image generation, 2023. URL: https://arxiv.org/abs/2308.00755.

[2] A. G. Buda, G. Coraglia, F. A. Genco, C. Manganini, G. Primiero, Bias amplification chains in ML-based systems with an application to credit scoring, in: Proceedings of the International Workshop on Bias, Risk, Explainability and the role of Logic and Logic Programming (BEWARE 2023), CEUR-WS, 2023.

[3] Y. Hirota, Y. Nakashima, N. Garcia, Quantifying societal bias amplification in image captioning, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13440–13449.

[4] S. L. Blodgett, et al., Bias and fairness in large language models: A survey, Computational Linguistics (2024).

[5] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, Bias in large language models: Origin, evaluation, and mitigation, 2024. URL: https://arxiv.org/abs/2411.10915.

[6] F. Haque, D. Xu, X. Niu, A comprehensive survey on bias and fairness in large language models, in: S. Yuan, F. Malliaros, X. Zheng (Eds.), Trends and Applications in Knowledge Discovery and Data Mining, Springer Nature Singapore, Singapore, 2025, pp. 83–101.

[7] H. Zhou, Z. Feng, Z. Zhu, J. Qian, K. Mao, Unibias: unveiling and mitigating LLM bias through internal attention and ffn manipulation, in: Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS 2024), Curran Associates Inc., Red Hook, NY, USA, 2024.

[8] X. Cheng, R. Chen, H. Zan, Y. Jia, M. Peng, Biasfilter: An inference-time debiasing framework for large language models, 2025. URL: https://arxiv.org/abs/2505.23829.
[9] L. Monteiro Paes, et al., Direct steering optimization for bias mitigation, arXiv preprint arXiv:2512.15926 (2025).
[10] Z. Siddique, I. Khalid, et al., Shifting perspectives: Steering vector ensembles for robust bias mitigation in llms, arXiv preprint arXiv:2503.05371 (2025).
[11] K. Kiashemshaki, M. J. Torkamani, N. Mahmoudi, M. S. Bilehsavar, Simulating a bias mitigation scenario in large language models (2025). URL: https://arxiv.org/abs/2509.14438. arXiv:2509.14438.
[12] S. Han, S. Avestimehr, C. He, Bridging the safety gap: A guardrail pipeline for trustworthy llm inferences, 2025. arXiv:2502.08142.
[13] S. G. Ayyamperumal, L. Ge, Current state of LLM risks and AI guardrails (2024). URL: https://arxiv.org/abs/2406.12934. arXiv:2406.12934.
[14] A. A. Syed, Guardrails for large language models: A review of techniques and challenges, URF Journals (2025).
[15] L. Lin, L. Wang, J. Guo, K.-F. Wong, Investigating bias in llm-based bias detection: Disparities between llms and human perception, in: Proceedings of COLING, 2025.
[16] M. Antonelli, L. Ceragioli, A. Buda, G. Primiero, A linear temporal logic of frequencies on series of events, 2026. URL: https://arxiv.org/abs/2604.10669.
[17] B. Legastelois, M.-J. Lesot, A. R. d'Allonnes, Typology of axioms for a weighted modal logic, International Journal of Approximate Reasoning 90 (2017) 341–358. URL: https://www.sciencedirect.com/science/article/pii/S0888613X17303997. doi:https://doi.org/10.1016/j.ijar.2017.06.011.
[18] A. G. Buda, G. Primiero, A logic for using information, Logique Et Analyse 265 (2025) 59–103.
[19] T. Chen, G. Primiero, F. Raimondi, N. Rungta, A computationally grounded, weighted doxastic logic, Stud Logica 104 (2016) 679–703. URL: https://doi.org/10.1007/s11225-015-9621-4. doi:10.1007/S11225-015-9621-4.
[20] N. J. Nilsson, Probabilistic logic, Artificial Intelligence 28 (1986) 71–87.
[21] F. Bacchus, Representing and Reasoning with Probabilistic Knowledge A Logical Approach to Probabilities, The MIT Press, 1990.
[22] R. Fagin, J. Halpern, N. Megiddo, A logic for reasoning about probabilities, Information and Computation 87 (1990) 78–128.
[23] K. W. Wagner, The complexity of combinatorial problems with succinct input representation, Acta informatica 23 (1986) 325–356.
[24] M. Antonelli, U. Dal Lago, P. Pistone, Towards logical foundations for probabilistic computation, Annals of Pure and Applied Logic 175 (2024).
[25] M. Antonelli, U. Dal Lago, P. Pistone, On counting propositional logic and Wagner's hierarchy, Theoretical Computer Science 966-967 (2023).
[26] A. Moskowakis, On a generalization of quantifiers, Fundamenta Mathematicae (1957).
[27] J. Kontinen, A logical characterization of the counting hierarchy, ACM Trans. Comput. Log. 10 (2009) 7:1–7:21.
[28] E. Kubyshkina, G. Primiero, A possible worlds semantics for trustworthy non-deterministic computations, Int. J. Approx. Reason. 172 (2024) 109212. URL: https://doi.org/10.1016/j.ijar.2024.109212. doi:10.1016/J.IJAR.2024.109212.
[29] F. A. D'Asaro, F. A. Genco, G. Primiero, Checking trustworthiness of probabilistic computations in a typed natural deduction system, J. Log. Comput. 35 (2025). URL: https://doi.org/10.1093/logcom/exaf003. doi:10.1093/LOGCOM/EXAF003.
[30] L. Ceragioli, G. Primiero, Trustworthiness preservation by copies of machine learning systems, Int. J. Approx. Reason. 192 (2026) 109638. URL: https://doi.org/10.1016/j.ijar.2026.109638. doi:10.1016/J.IJAR.2026.109638.
[31] L. Hutschenreiter, C. Baier, J. Klein, Parametric markov chains: Pctl complexity and fraction-free gaussian elimination, Electronic Proceedings in Theoretical Computer Science 256 (2017) 16–30. URL: http://dx.doi.org/10.4204/EPTCS.256.2. doi:10.4204/eptcs.256.2.

[32] N. Jansen, S. Junges, J.-P. Katoen, Parameter Synthesis in Markov Models: A Gentle Survey, Springer Nature Switzerland, Cham, 2022, pp. 407–437.

[33] M. Maaz, T. C. Y. Chan, Formal verification of markov processes with learned parameters, 2025. URL: https://arxiv.org/abs/2501.15767. arXiv:2501.15767.

[34] G. Coraglia, F. A. D'Asaro, F. A. Genco, D. Giannuzzi, D. Posillipo, G. Primiero, C. Quaggio, Brioxalkemy: a bias detecting tool, in: G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, G. Primiero (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), Rome, Italy, November 6, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023, pp. 44–60. URL: https://ceur-ws.org/Vol-3615/paper4.pdf.

[35] A. G. Buda, G. Coraglia, F. A. Genco, C. Manganini, G. Primiero, Bias amplification chains in ml-based systems with an application to credit scoring, in: G. Coraglia, F. A. D'Asaro, A. Dyoub, F. A. Lisi, G. Primiero (Eds.), Proceedings of the 3rd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), Bolzano, Italy, November 26, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 77–86. URL: https://ceur-ws.org/Vol-3881/paper9.pdf.

[36] A. G. Buda, G. Coraglia, F. A. Genco, C. Manganini, G. Primiero, Assessing the risk of discrimination with BRIO: A use case from the financial sector, in: D. Pedreschi, M. Milano, I. Tiddi, S. Russell, C. Boldrini, L. Pappalardo, A. Passerini, S. Wang (Eds.), Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence, HHAI 2025, Pisa, Italy, June 9-13, 2025, Frontiers in Artificial Intelligence and Applications, IOS Press, 2025, pp. 483–485. URL: https://doi.org/10.3233/FAIA250668. doi:10.3233/FAIA250668.