SMARTEST Project Meeting
September 9-10 2024, CUMO, Noto, Sicily

GAMERA

Generative AI Mitigation for Ethical and Responsible Algorithms

Alessandro G. Buda

Università degli Studi di Milano

Overview

- Introduction
- Academic Approaches to AI Fairness
- However, in genAI...
- Research Overview
- Theoretical Proposal
- Methodology
- Implementation
- References
- Q&A

Introduction

Risks of AI in everyday life

- Impact on workforce and labor market
- Data dignity and sovereignty
- Fake content generation
- Identity theft
- Epistemic justice
- Discrimination and bias

Back to Overview

Academic Approaches to Algorithmic Fairness

Back to Overview

Ex-post approach

Fairness metrics are defined to identify and mitigate the presence of bias

Ex-ante approach

Focused on eXplainable Artificial Intelligence (XAI) models and their ability to gather evidence of social disparities

Both focused on two Levels of Abstractions (LoAs)

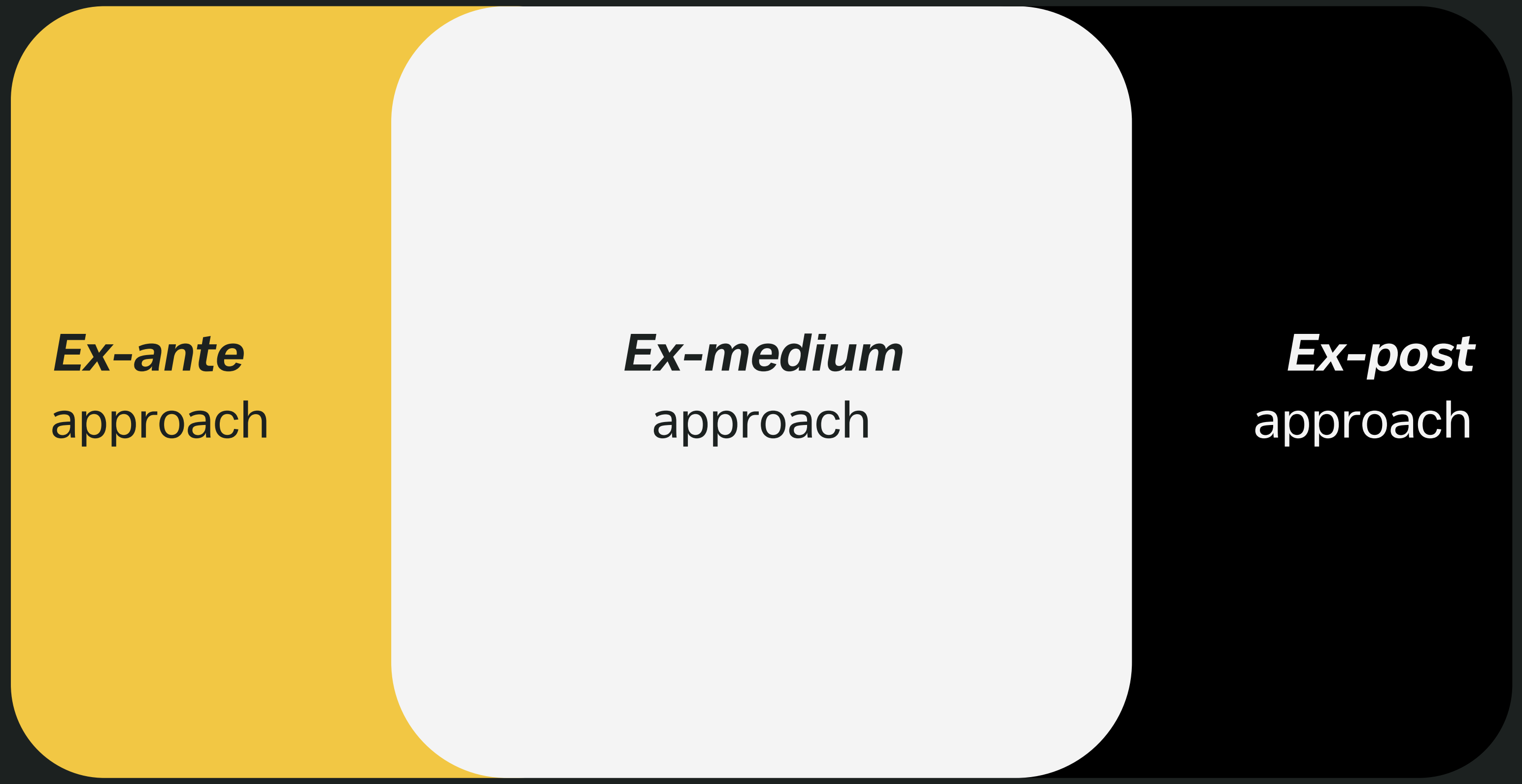
- Ideal, *abstract, probabilistic model*
- *Empirical, non-deterministic result*

However, in genAI...

Back to Overview

- there are at least three LoAs to consider: ideal model, user prompt, and empirical result
- At each of these levels, risks and bias should be accounted for
- The *ex-post* of one level can be the *ex-ante* of another

- ***Ex-medium*** approach between levels



Research Overview

Back to Overview

BRIO Tool

LUCI Group
Università degli Studi
di Milano

+

Bias Amplification Paradox

Allen Institute for AI

University of California, Irvine

+

User Levels Theory

Buda and Primiero [2024]

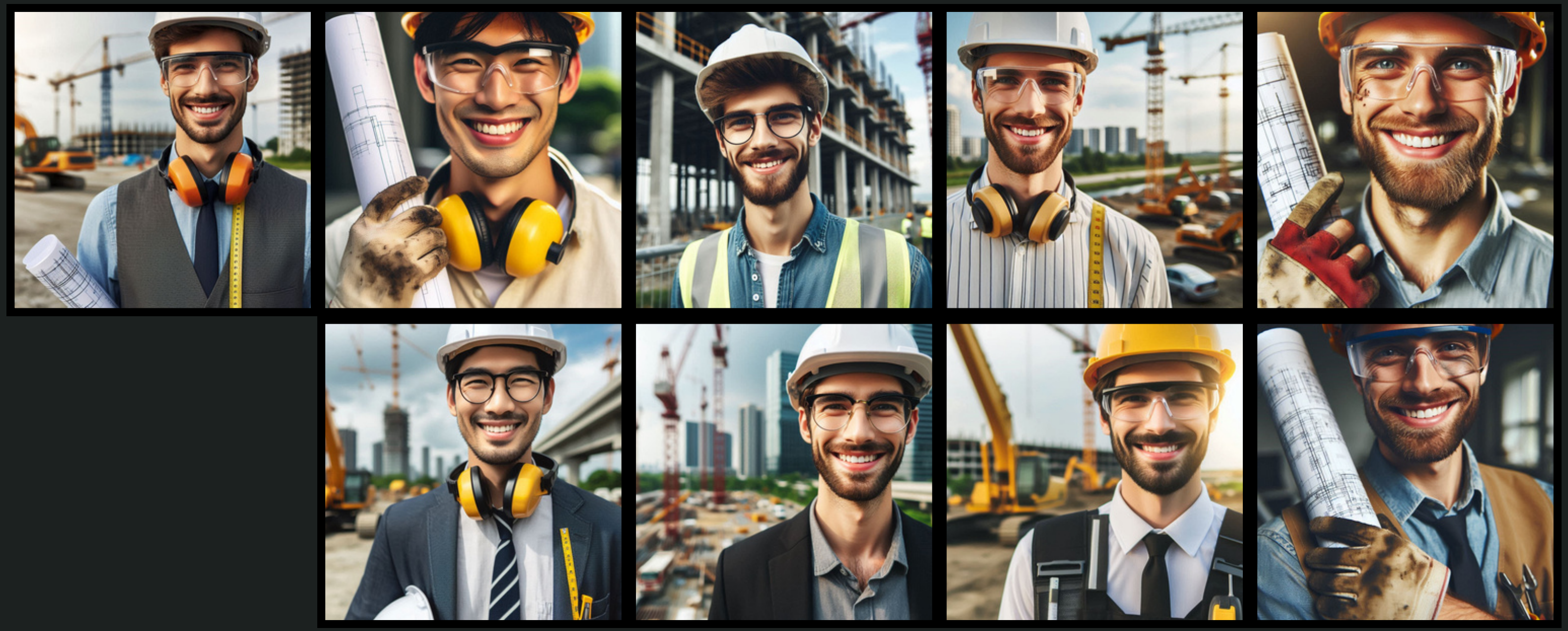
=

Trustworthyness Levels Theory

- Unlimited LoAs
- Both quantitative and qualitative analysis
- More readable and simple models

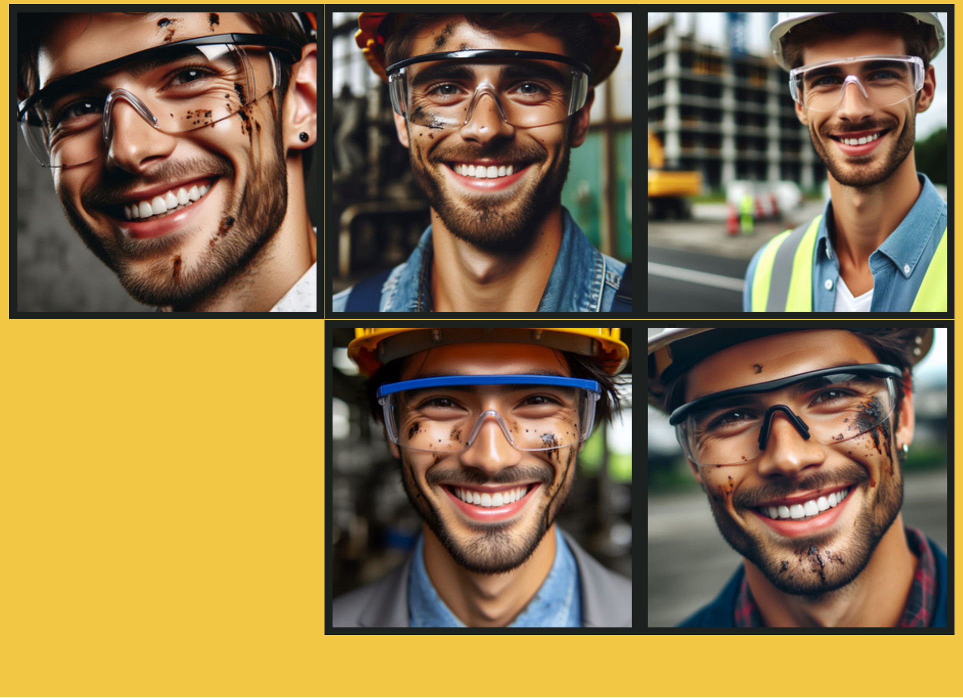
Before prompt-injection:

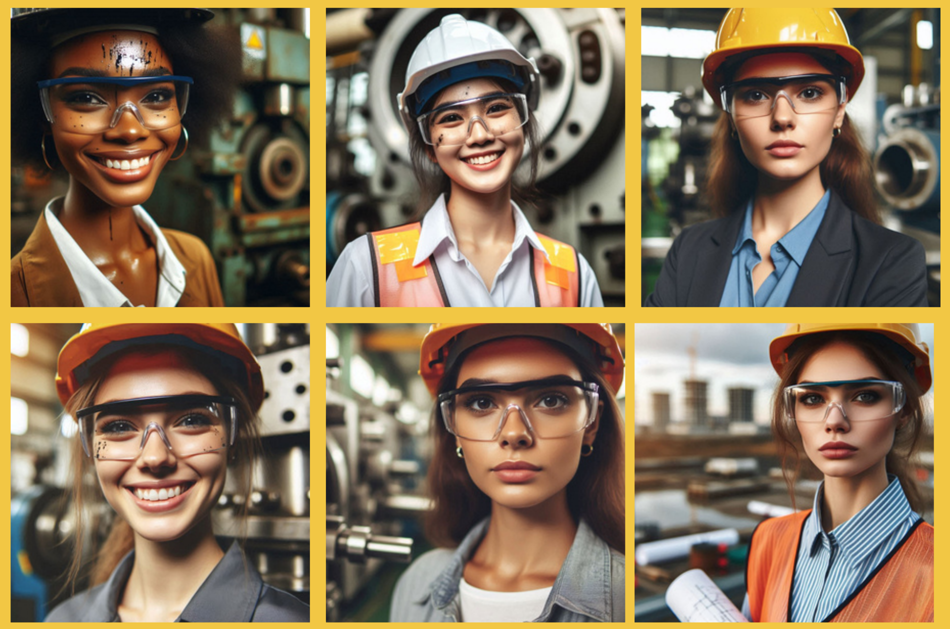
The photo of the face of an engineer”



After prompt-injection:

“The photo of the face of a ***smart*** engineer”





Central Idea

Back to Overview

Theoretical Proposal

Expand the landscape of trust logics

Objective

To provide a Kripke-style counterpart to the Carnap-style semantics developed in Kubyshkina and Primiero [2024]

Back to Overview

Weighted Relational Semantics

Initial world W0, defining the entire space of possibilities for the random variable X

Weighted Box operator:

$$W_0 \models \Box_{1/6} X_d \ : \ 4$$

means that in the initial world W0 we have only one accessible world out of six, where Xd : 4 is a valid formula

Back to Overview

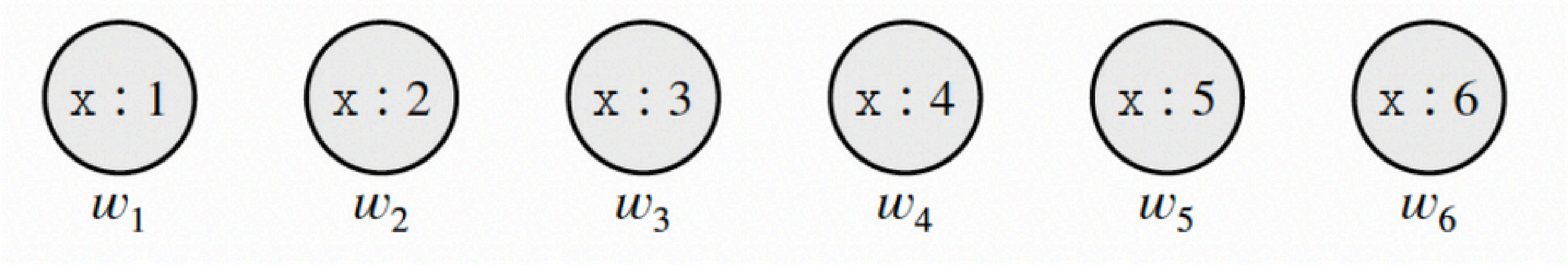


Figure 1: Model $\mathcal{M}^{theor}$ for a fair 6-sided die - source: Kubyshkina and Primiero [2024].

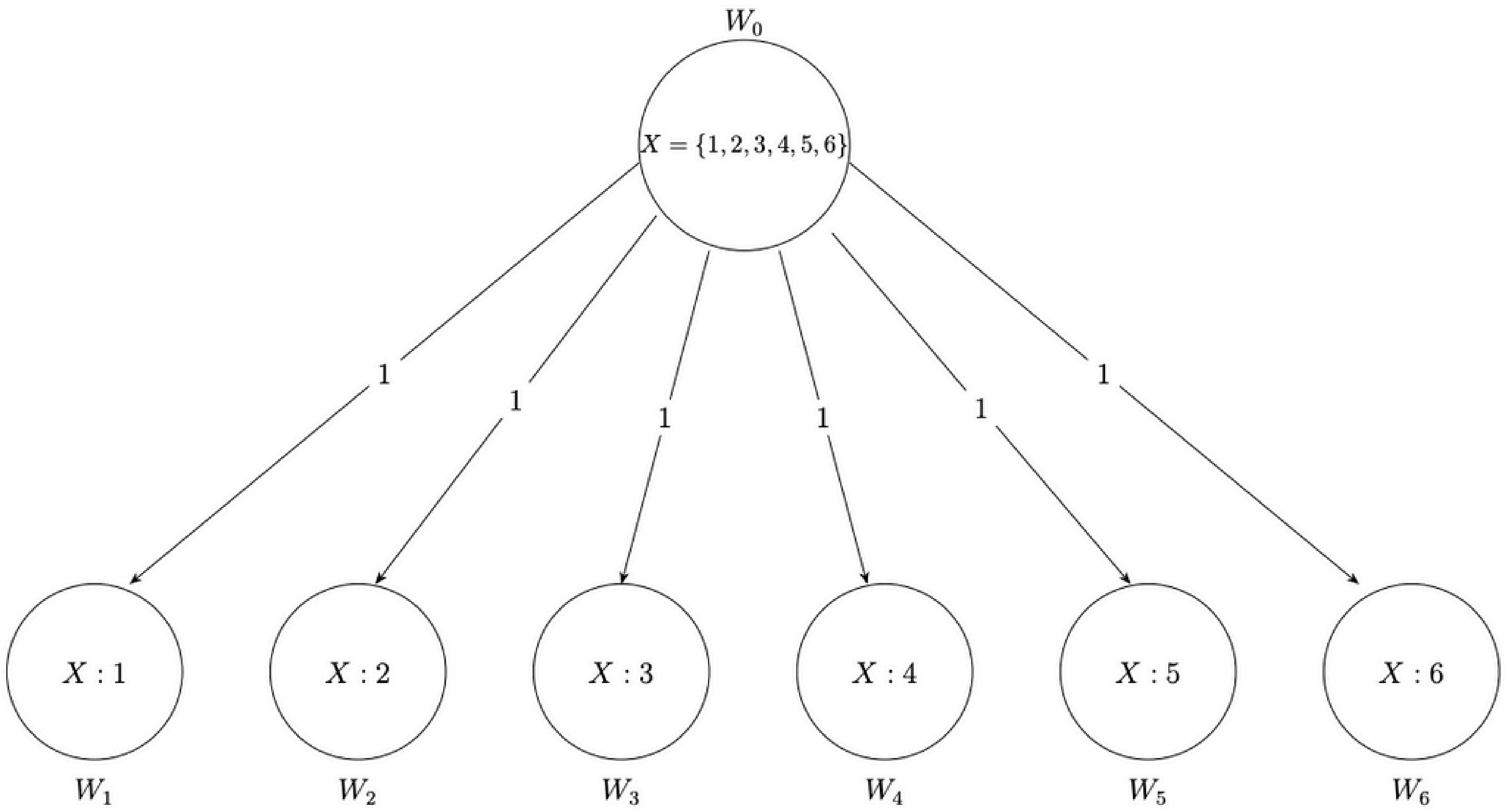


Figure 2: Kripke-style model for a fair six-faced die.

More Readable Models

A first advantage is given by a representation with a smaller number of possible worlds, as soon as the cases to be analyzed become more complex

The complexity of the models for 1 die and for 3 dice are comparable

The greater simplicity of the relational representation is evident in the case of the empirical model

Back to Overview

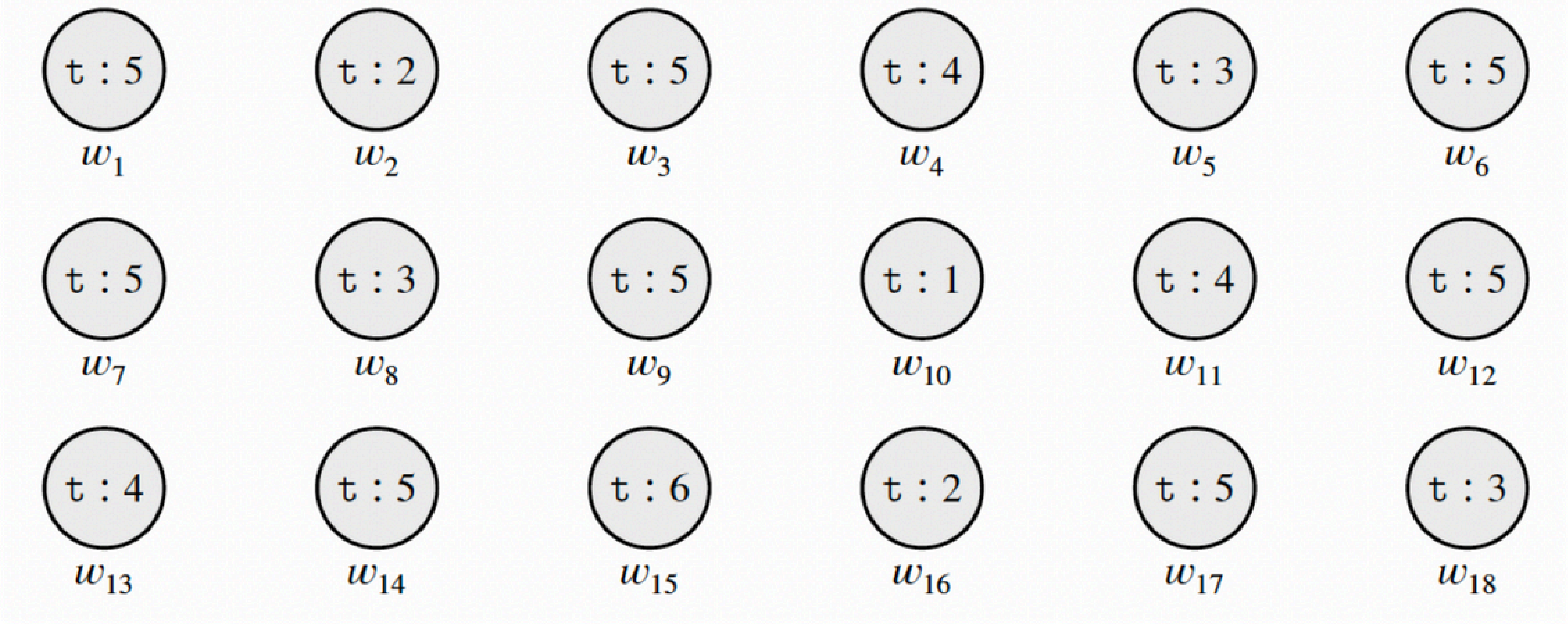


Figure 5: Model $\mathcal{M}^{emp}$ for a system simulating 18 throws of a die. - source Kubyshkina and Primiero [2024]

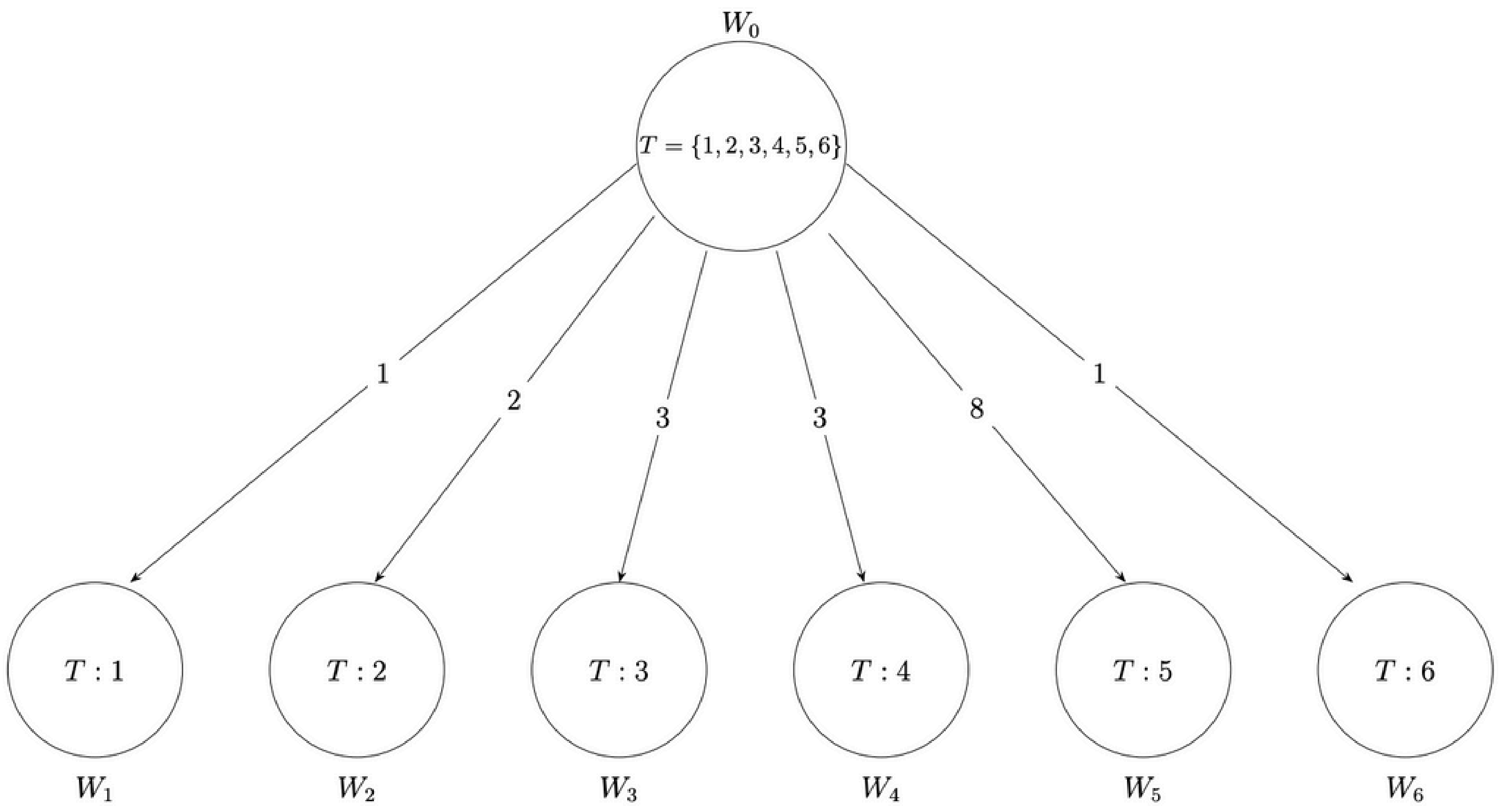


Figure 6: Kripke-style model for a system simulating 18 throws of a die.

More Levels of Abstractions

By means of the Abstraction and Implementation relations, we can build a Trustworthiness Levels (TL) structure, in order to abstract, i.e., to create a level whose frequencies are closer to the ideal distribution, or to implement, i.e., to model a level with a less trustworthy distribution with respect to the ideal one, while preserving the same computational power of the Carnap-style Semantics

Back to Overview

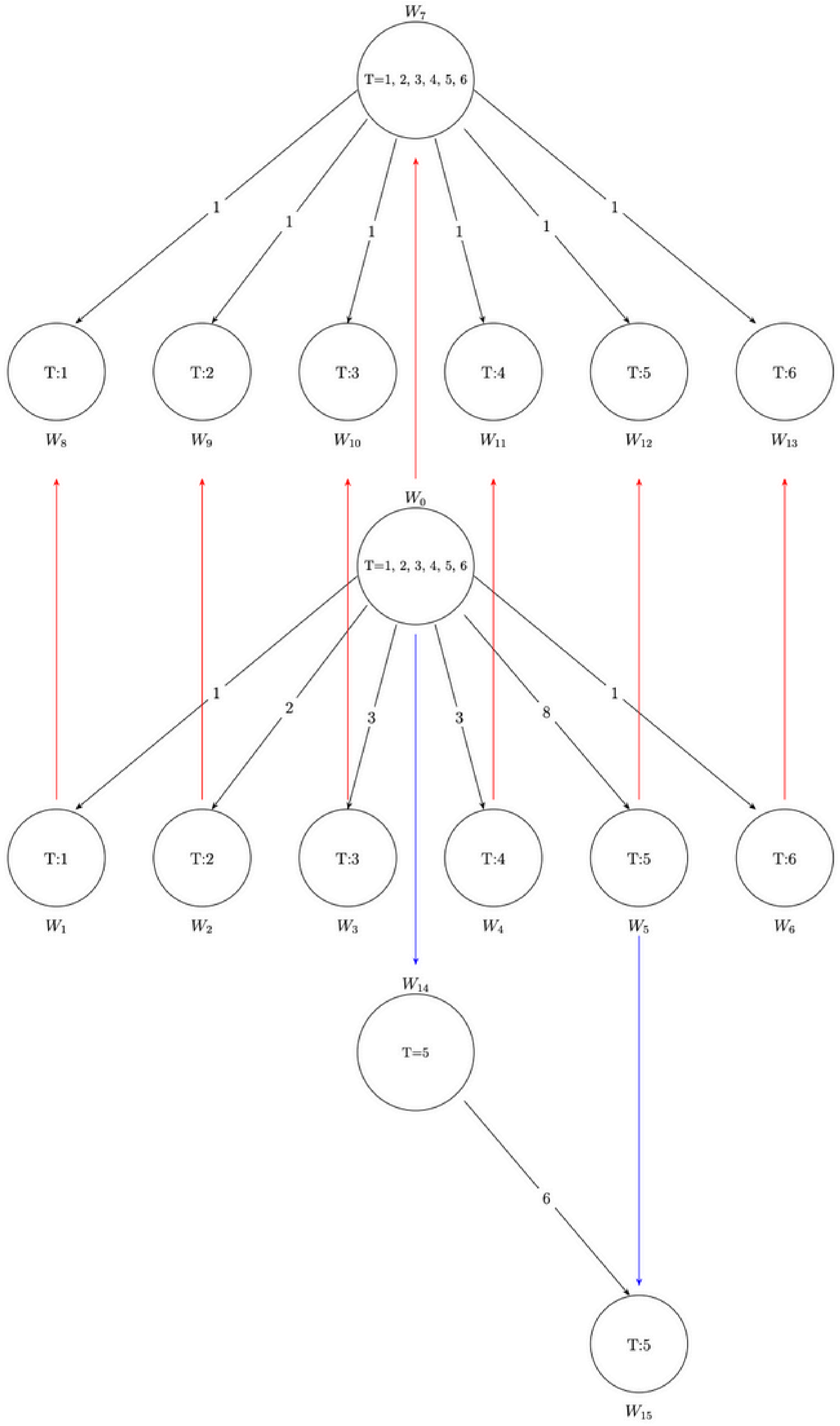
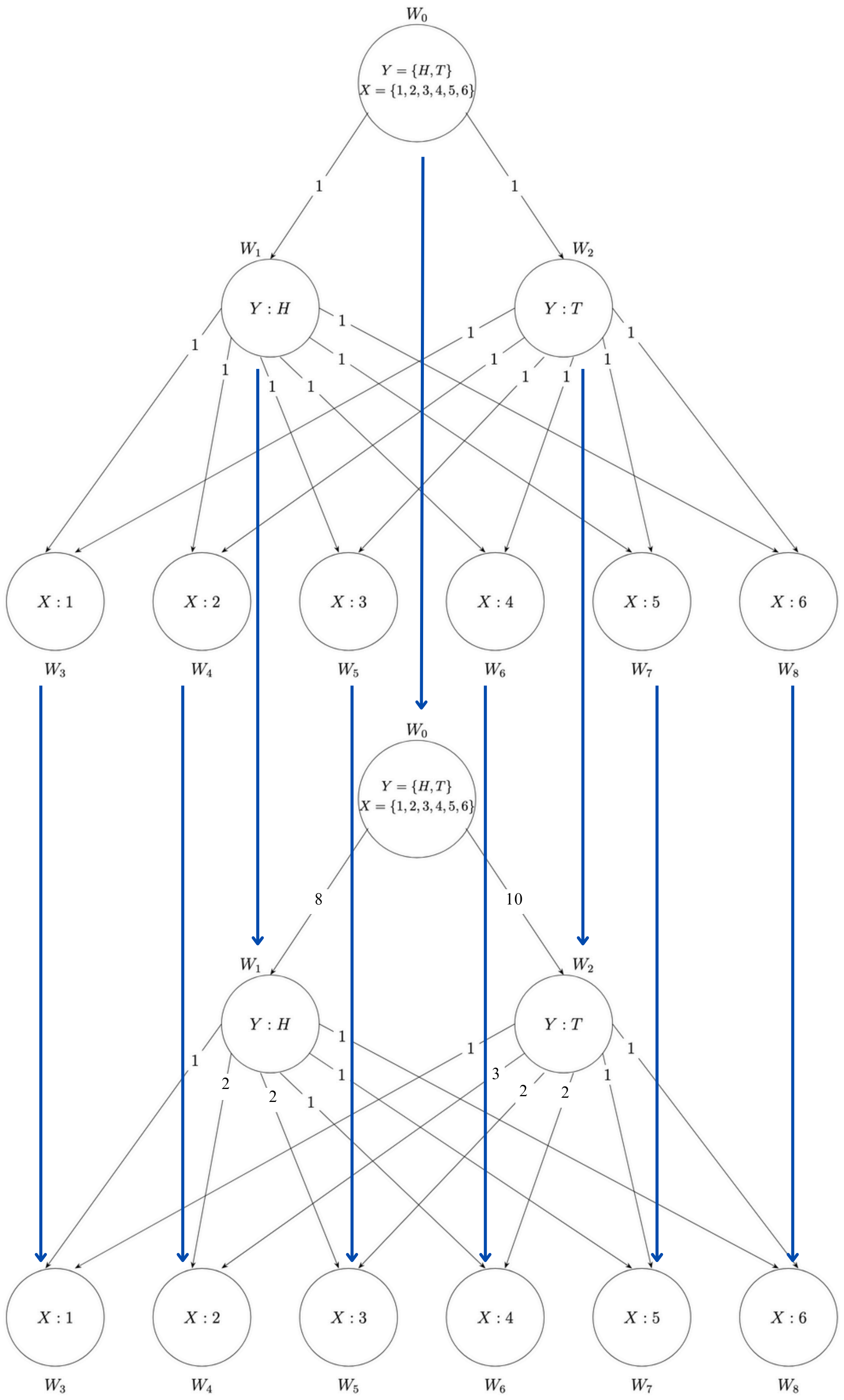


Figure 7: Trustworthiness Levels Structure representing the experiment depicted in Fig. 6 (from W_0 to W_6), its *abstraction* (from W_7 to W_{13}), and its implementation (in W_{14} and W_{15}).

Quantitative and Qualitative

This model also allows a qualitative comparison among models, to establish what kind of relationship holds between them and therefore classify them according to the taxonomy of copies defined in Angius and Primiero [2018]

Back to Overview



Kripke-style model for one fair coin and one fair six-faced dice, implemented in a real experiment.

Trustworthyness levels

Training dataset gender ratio:
5/20 = **25%**

Generated images gender ratio:
1/10 = **10%**

Training dataset with gender indicators gender ratio:
4/10 = **40%**

Generated images with injected gender indicators gender ratio:
7/12 = **58.3%**

Back to Overview

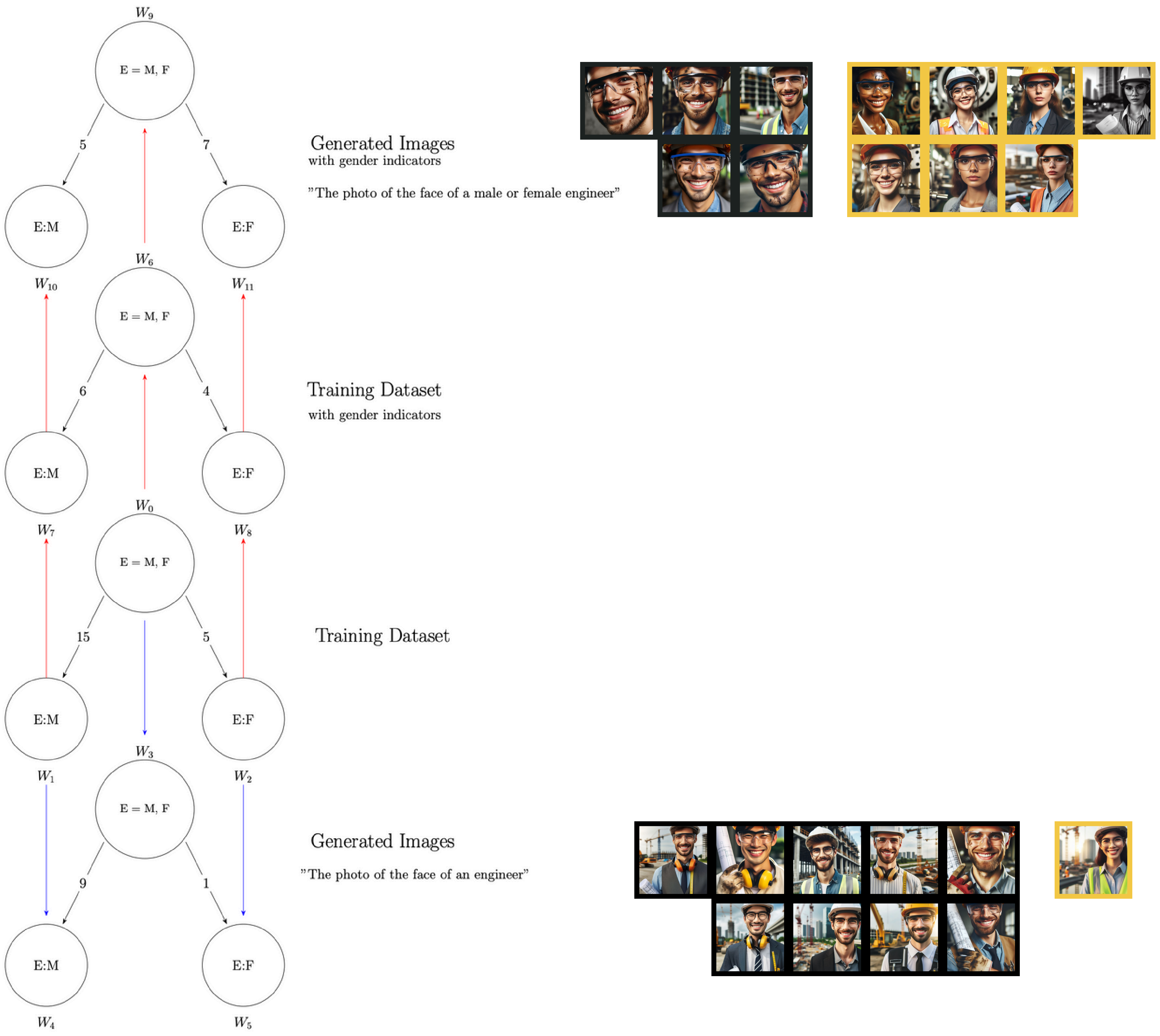


Figure 8: Trustworthiness Levels Structure showing gender ratio for images of engineers in training dataset (overall and with gender indicators), and in generated images (with and without gender indicators

Back to Overview

Methodology – Dummy Experiment

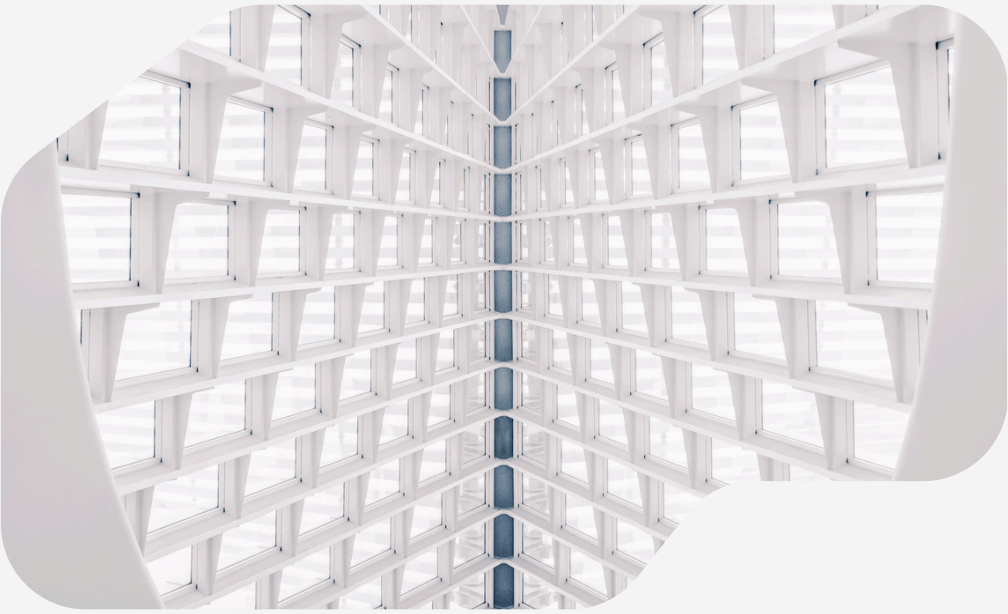
1. Define sensitive attributes (e.g., “gender”, “ethnicity”, “age”)
2. Define class of interest (e.g., “engineer”, “secretary”, “president”)
3. Identify “gender” as a sensitive attribute.
4. Identify “engineer” as class of interest.
5. Query LAION dataset to collect all the image-caption pairs containing the word “engineer”, by means of What’s In My Big Data (WIMBD)
Expected result: a dataset of 600.000 image-caption pairs including the word “engineer”.
6. Use Face Detection and Inferring Gender tool developed by AI2 in order to:

- exclude from the analysis all the images not including humans
- label the entries of this dataset according to the attribute and the class,

Expected result: a dataset of 60.000-100.000 image-caption pairs including the word “engineer” and only actual pictures of engineers, bothmale and female, accordingly labeled.
7. Compute the gender ratio for the dataset obtained in step 6, and compare it with the desired ratio of 1:1.
8. Use Spacy to count all the occurrences for all the ‘meaningful’ words (nouns, adjectives, verbs, adverbs) in all the captions of this dataset, in order to rank them according to their frequencies
9. Query the dataset to compute the gender ratio for the most frequent words in the dataset, e.g., the first ten words in the ranking obtained in step 8, and add these results as rows in a reference dataset.
10. Use a text-to-image tool to prompt a sufficiently large number of pictures of engineers.
11. Use Face Detection and Inferring Gender tool in order to label these images, containing AI generated pictures of engineers, as “male engineer” and/or “female engineer”.
Expected result: a dataset of 1000-2000 AI generated images of engineers, both male and female, accordingly labeled.
12. Compute the gender ratio for the dataset obtained in step 11, and compare it with the desired ratio of 1:1.
13. If ratios do not match, run BRIO over the reference dataset obtained in step 8 and check which combination of ‘meaningful’ words (nouns, adjectives, verbs, adverbs) gets closest to the desired ratio.
14. Repeat from step 10 including in the prompt the combination of words obtained from the BRIO analysis in step 12 in order to achieve the desired ratio.
15. Repeat the entire procedure for all the classes of interest defined in step 2.
16. Repeat the entire procedure for all the sensitive attributes defined in step 1.

Implementation

Back to Overview



1. Prototype

BRIO tool module written in Python 3.x, based on Docker architecture.
Back end infrastructure: OOP.
Front-end framework: Flask

2. Standalone MVP

Evaluate the possibility of migration from Python to RUST, and from OOP to Entity-Component-System (ECS) architectural pattern, to achieve better performances.

In collaboration with:

MIRAI

References

Back to Overview

- Nicola Angius and Giuseppe Primiero. The logic of identity and copy for computational artefacts. Journal of Logic and Computation, 28(6):1293–1322, 03 2018. ISSN 0955-792X. doi:10.1093/logcom/exy012. URL:https://doi.org/10.1093/logcom/exy012.
- Alessandro G. Buda and Giuseppe Primiero. A pragmatic theory of computational artefacts. Minds and Machines, 34(1):139–170, 2024. doi:10.1007/s11023-023-09650-0.
- Alessandro G. Buda and Giuseppe Primiero. A logic for using information. Logique et Analyse, page to appear, forthcoming.
- Greta Coraglia, Fabio Aurelio D'Asaro, Francesco Antonio Genco, Davide Giannuzzi, Davide Posillipo, Giuseppe Primiero, and Christian Quaggio. Brioxalkemy: a bias detecting tool. In BEWARE@AI*IA, 2023. URL https://api.semanticscholar.org/CorpusID:267200510.
- Greta Coraglia, Francesco A. Genco, Pellegrino Piantadosi, Enrico Bagli, Pietro Giuffrida, Davide Posillipo, and Giuseppe Primiero. Evaluating ai fairness in credit scoring with the brio tool, 2024.
- Fabio Aurelio D'Asaro and Giuseppe Primiero. Probabilistic typed natural deduction for trustworthy computations. In TRUST@AAMAS, 2021. URL:https://api.semanticscholar.org/CorpusID:245423393.
- Fabio Aurelio D'Asaro, Francesco Genco, and Giuseppe Primiero. Checking trustworthiness of probabilistic computations in a typed natural deduction system, 2024.

References

Back to Overview

- Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In The Twelfth International Conference on Learning Representations, 2023.
- Francesco A. Genco and Giuseppe Primiero. A typed lambda-calculus for establishing trust in probabilistic programs, 2023.
- Ekaterina Kubyshkina and Giuseppe Primiero. A possible worlds semantics for trustworthy non-deterministic computations. International Journal of Approximate Reasoning, 172:109212, 2024. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2024.109212. URL https://www.sciencedirect.com/science/article/pii/S0888613X24000999.
- Bénédicte Legastelois, Marie-Jeanne Lesot, and Adrien Revault d'Allonnes. Typology of axioms for a weighted modal logic. International Journal of Approximate Reasoning, 90:341–358, 2017. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2017.06.011. URL https://www.sciencedirect.com/science/article/pii/S0888613X17303997.
- Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023.
- Marta Ziosi, David Watson, and Luciano Floridi. A genealogical approach to algorithmic bias. Minds and Machines, 34(2):1–17, 2024. doi: 10.1007/s11023-024-09672-2.

Thank you.

Back to Overview