

Trustworthy Copies of Non-deterministic Computations

Giuseppe Primiero



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FILOSOFIA
"PIERO MARTINETTI"



Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

Checking Trustworthiness of Copies

Conclusions

Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

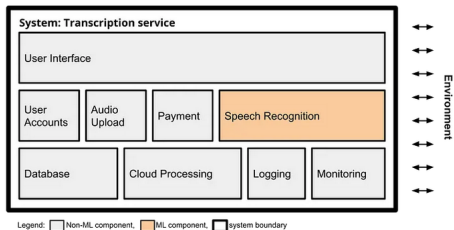
Checking Trustworthiness of Copies

Conclusions

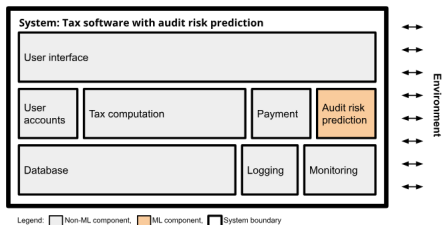
The Background Issue

- ▶ Consider a non-deterministic computational artefact of the ML family integrated as a component in an industrial setting
- ▶ Digital twins of artefacts integrating ML processes may diverge wrt to the class of outputs they produce based either on training or on model tuning.

Some examples, [Kästner, 2022]



Architecture sketch of a transcription system, illustrating the central ML component for speech recognition and many non-ML components.



An architecture sketch of the tax system, illustrating the ML component for audit risk as an addition to many non-ML components in the system.

Which Notion of Copy



1. Under no knowledge of the training dataset for the copied model, the same trained model is used on a new input dataset
2. Under no knowledge of the copied model, a new model is trained on the same dataset

The Background Issue (simplified)



- ▶ An ML system may be evaluated for trustworthiness for protected attributes wrt to given criteria
- ▶ The outputs of any copy (in the above sense) for the same attribute may be equivalent to each subclass of possible events (with appropriate probabilities) of the original system

A toy example

Consider a GenAI system for text-to-image generation and the distribution of some property of interest in its output (e.g. gender).

For example, the system on the prompt

Smart engineer

executed n times may output on average $3/4$ of male engineers, and $1/4$ of female engineers. It may be the case that on the population of interest this appears as a biased output. This behaviour can be verified.

Our Question

Assume now that a new system is trained based on the same data. Or the same trained model is used on new data. We understand the latter as a copy of the former system under no knowledge of the underlying model or of the training data.

Our Question

Assume now that a new system is trained based on the same data. Or the same trained model is used on new data. We understand the latter as a copy of the former system under no knowledge of the underlying model or of the training data.

Question

How does one establish to which degree (i.e. wrt which class of outputs) such a copy is safe in order to evaluate its trustworthiness?

The formal aim

Task

Formalize trustworthiness evaluation of copies based on how they respect validity conditions wrt the original system.

Our Strategy

- ▶ Start from a(n existing) formal trustworthiness evaluation
- ▶ Use (already available) validity conditions for ML systems
- ▶ Construct different notions of trustworthiness associated to the various validity conditions

Why is this useful

- ▶ In absence of knowledge of the training set, useful to assess whether the copy and the model may have similarly structured trained models
- ▶ In absence of knowledge of the model, useful to assess bias amplification/reduction of the copy wrt the model
- ▶ In both cases, anticipate safety/liveness properties of the copied system (technically: through bisimulation of progress and termination results)

Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

Checking Trustworthiness of Copies

Conclusions

The single output trustworthiness evaluation, [D'Asaro et al., 2024]

Measure of trustworthiness as distance of observed behaviour against expected behaviour:

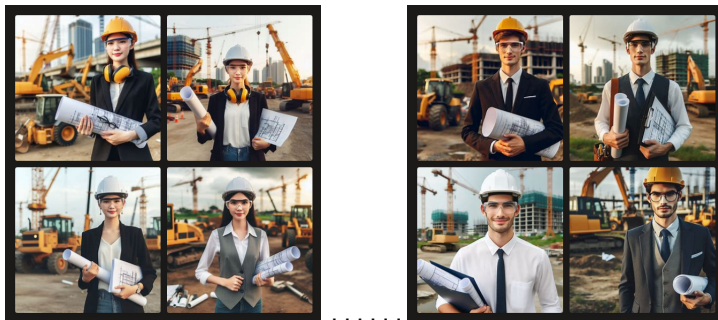
Given a random variable x with output α with probability a under a distribution Γ , evaluate the related trustworthiness of process u under a possibly unknown distribution Δ by analysing the frequency f of output α in n runs by assessing how much f diverges from probability a parametric wrt to n .

The single output trustworthiness evaluation

$$\frac{\Gamma \vdash x : \alpha_a \quad \Delta \vdash u_n : \alpha_f \quad |a - f| \leq \epsilon(n)}{\Gamma, \Delta \vdash \text{Trust}(u_n : \alpha_f)} \text{IT}$$

With Γ a known distribution of reference, Δ a possibly unknown distribution of properties in the training set and α the protected attribute of interest.

The toy example



How fair does it appear the male/female ratio wrt to the ideal frequency in a given population?

Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

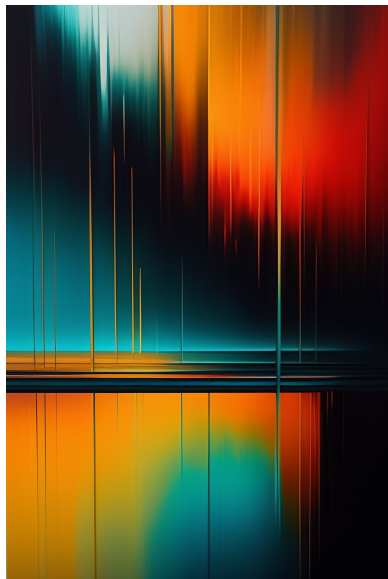
Checking Trustworthiness of Copies

Conclusions

Analysing the behaviour of digital copies

- ▶ The behaviour of copies may be considered in the light of validity conditions in experimental simulative sciences
- ▶ In this context one considers a simulation system (i.e. a copy) and the validity of its outputs against the output of its model
- ▶ In this case, the comparison is not the ideal (as in simulative sciences) but the observed behavior of the copied system
- ▶ The associated notion of validity has been labeled *experimental* in [Primiero, 2020]

Validity Principles [Manganini and Primiero, 2024]



- ▶ An analysis of validity conditions for ML systems (as non-deterministic computational systems) defined over probabilities for training and test data of the model on the one hand, and test data and features of the copied system on the other
- ▶ different simulation relations may be defined over the two observed systems, and accordingly different validity conditions may be extracted.

Validity Principles [Manganini and Primiero, 2024]

Definition (Justifiably Valid Model)

A machine learning based computational model is justifiably valid for a given target system only if at least a weak probabilistic bisimulation relation holds between them, i.e.:

- ▶ the test data satisfy **all and only** probabilistic assignments as the training data, and
- ▶ the assignments of the machine learning model to data points in the test data match **only** features in the target system.

Validity Principles [Manganini and Primiero, 2024]

Definition (Weakly Valid Model)

A machine learning based computational model is weakly valid for a given target system only if at most a weak probabilistic simulation relation holds between them, i.e.:

- ▶ the test data satisfy probabilistic assignments with **at least** the same values as in the training data, and
- ▶ the assignments of the computational model for features to points in the test data are **at least** the same as in the target system.

Validity Principles [Manganini and Primiero, 2024]

Definition (Almost Valid Model)

A machine learning based computational model is almost valid for a given target system only if at most an approximate probabilistic simulation relation holds between them, i.e.:

- ▶ the test data satisfy probabilistic assignments with at least the same values as the training data, **but not only**, and
- ▶ the assignments of the computational model for features to points in the test data are at least the same as in the target system, **but not only**.

Formalising Validity Principles

Question

How do we translate formally these criteria for trustworthiness verification?

Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

Checking Trustworthiness of Copies

Conclusions

The artefact

- ▶ Consider a non deterministic process t , valid under a possibly opaque given probability distribution Γ of features, with outputs $\alpha^1, \dots, \alpha^n$,
- ▶ execute it n times, each output occurs with frequency f^1, \dots, f^n respectively

$$\Gamma \vdash t_n : \alpha_{f_1, \dots, f_n}^{1, \dots, n}$$

The copy

- ▶ Consider a copy u , valid under a given probability distribution Δ of features, with outputs $\alpha^1, \dots, \alpha^n$,
- ▶ execute it n times, each output occurs with frequency g_1, \dots, g_n respectively

$$\Gamma \vdash u_n : \alpha_{g_1, \dots, g_n}^{1, \dots, n}$$

The trustworthiness question

Question

Is u a trustworthy copy of t ?

The trustworthiness question

Question

Is u a trustworthy copy of t ?

We aim at answering by considering which notion of validity does u satisfy wrt t .

A toy example: weak bisimulation

Consider a fair coin

$$\{c_n : H_{1/2}, c_n : T_{1/2}\}$$

and a perfect digital copy

$$\{d_n : H_{1/2}, d_n : T_{1/2}\}$$

Justifiably Trustworthy Copy

$$\frac{\Gamma \vdash t : \alpha_{f_1}^1, \dots, \Gamma \vdash t : \alpha_{f_n}^n \quad \Delta \vdash u_n : \alpha_{g_1}^1, \dots, \Delta \vdash u_n : \alpha_{g_n}^n}{\Gamma, \Delta \vdash JTrust(u_n : \alpha^1, \dots, \alpha^n)} \text{IT}$$

where $\forall \alpha^1 \dots \alpha^n \mid f_i = g_i$

A toy example: weak simulation

Consider a fair coin

$$\{c_n : H_{1/2}, c_n : T_{1/2}\}$$

and a digital copy

$$\{d_n : H_{1/3}, d_n : T_{2/3}\}$$

Weakly Trustworthy Copy

$$\frac{\Gamma \vdash t : \alpha_{f_1}^1, \dots, \Gamma \vdash t : \alpha_{f_n}^n \quad \Delta \vdash u_n : \alpha_{g_1}^1, \dots, \Delta \vdash u_n : \alpha_{g_n}^n}{\Gamma, \Delta \vdash WTrust(u_n : \alpha^1, \dots, \alpha^n)} \text{IT}$$

where $\exists \alpha^i \mid g_i \geq f^i$

A toy example: approximate simulation

Consider a fair die

$$\{d_n : 1_{1/6}, \dots, d_n : 6_{1/6}\}$$

and a digital copy

$$\{c_n : 1_{1/6}, \dots, c_n : 7_{1/6}\}$$

Almost Trustworthy Copy

$$\frac{\Gamma \vdash t : \alpha_{f_1}^1, \dots, \Gamma \vdash t : \alpha_{f_n}^n \quad \Delta \vdash u_n : \alpha_{g_1}^1, \dots, \Delta \vdash u_n : \alpha_{g_n}^n}{\Gamma, \Delta \vdash ATrust(u_n : \alpha^1, \dots, \alpha^n)} \text{IT}$$

where $\exists \alpha^i \mid g_i > 0 = f^i$

Introduction

Preamble: Formal Verification of Trustworthiness

Validity Conditions for non-deterministic Computations

Checking Trustworthiness of Copies

Conclusions

Summary and next steps







- ▶ Formal verification of trustworthiness of digital copies
- ▶ Evaluating trust beyond a binary property
- ▶ Implementation on the BRIO platform by



Thanks

References I

-  D'Asaro, F. A., Genco, F., and Primiero, G. (2024).
Checking trustworthiness of probabilistic computations in a typed natural deduction system.
-  Kästner, C. (2022).
Machine Learning in Production: From Models to Products.
(a final copy will be published late 2024 or 2025 by MIT Press).
-  Manganini, C. and Primiero, G. (2024).
Philosophy of Science for Machine Learning, chapter Defining Formal Validity Criteria for Machine Learning Models.
Springer Nature.
-  Primiero, G. (2020).
On the Foundations of Computing.
Oxford University Press, Inc., USA.