TOWARDS A RELATIONAL SEMANTICS FOR EVALUATING TRUSTWORTHINESS

Ekaterina Kubyshkina

LUCI Group, University of Milan
(joint work with Giuseppe Primiero)

Abstract

As noted in the Ethics Guidelines for Trustworthy AI (2018), trustworthiness seems to be a prerequisite for people and societies to develop, deploy and use AI systems. The aim of our work is to provide a relational semantics to evaluate whether an AI system is trustworthy. Our starting point is the *Trustworthy Probabilistic Typed Natural Deduction* (TPTND) developed by D'Asaro & Primiero (2021, 2022), which makes trustworthiness checkable by combining typed natural deduction with probabilistic reasoning. Although the system TPTND has some intuitively desirable properties for capturing the notion of trustworthiness, no relational semantics was provided by the authors. We aim to fill this lacuna, thus relating this framework with mainstream works in modal logic.

System TPTND

The system TPTND is defined over the language containing expressions about processes and their possible outcomes. The following distinction between these expressions can be highlighted:

1. Expressions about idealized and real processes and their outcomes:
   - $x : \alpha$ means that a random variable x has value α;
   - $t : \alpha$ means that a process t produces an output α.
2. Expressions about the probability of a process to produce an output:
   - $x : \alpha_a$ means that the probability of a random variable x to have value α is a;
   - $t_n : \alpha_a$ means that, after n executions of a process t, an output α was produced with frequency a;
   - $t_n : \alpha_{\tilde{a}}$ means that a process t produces an output α with the expected probability $\tilde{a}$ over n executions.

TPTND contains four fragments:

- **distribution construction rules**, which define the contexts as list of assumptions on the probability distributions of processes to have certain outputs;
- **rules for random variables**, which define operations on the expected probability of random variables to produce outputs;
- **sampling rules**, which define operations on observed frequencies of processes to produce outputs and correlate them with the expected probabilities;
- **trust fragment**, which defines a procedure for the decision of whether a process is trustworthy.

Trust fragment

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

$$\frac{\Gamma \vdash Trust(u_n : \alpha_f)}{\Gamma, x_u : \alpha_{[a-\epsilon(n), a+\epsilon(n)]} \vdash u_n : \alpha_f}\text{ET}$$

$$\frac{\Gamma, x : \alpha_a :: distribution \quad \Delta \vdash u_n : \alpha_f \quad \mid a - f \mid > \epsilon(n)}{\Gamma, \Delta \vdash UTrust(u_n : \alpha_f)}\text{IUT}$$

$$\frac{\Gamma \vdash UTrust(u_n : \alpha_f)}{\Gamma, x_u : \alpha_{[0,1]-[a-\epsilon(n), a+\epsilon(n)]} \vdash u_n : \alpha_f}\text{EUT}$$

Semantics for TPTND

In what follows we consider trustworthiness of a process as hypothesis testing on the distance between the frequency of an observed output of this process and its intended probability. To capture this idea semantically, we consider two relational models: theoretical and empirical ones.
The **theoretical model** represents an ideal distribution of producing some output by a process, which is expected by the agent evaluating the trustworthiness of this process. These models characterize the distribution construction rules and the rules for random variables.
The **empirical model** represents chronologically ordered series of experiments. These models characterize the sampling rules which do not include statements involving random variables.
In order to characterize the sampling rules involving random variables and, most importantly the trust fragment, we take a **fusion of theoretical and empirical models**.

Theoretical models

Definition 1. *Let $\mathcal{M}^{theor}$ be $(W^{theor}, R^{theor}, v^{theor})$, such that W^{theor} is non-empty set of worlds $w_1, ..., w_n$ such that $w_1, ..., w_n$ are mutually exclusive and exhaustive sets of ES^x, $R^{theor} \subseteq W^{theor} \times W^{theor}$ is an equivalence relation, $v^{theor} : ES^x \to P(W)$ is a valuation function.*

1. *$\mathcal{M}^{theor}, w_i \models_t x : \alpha$ iff $w \in v(x : \alpha)$;*
2. *$\mathcal{M}^{theor}, w_i \models_t x : (\alpha + \beta)$ iff $\mathcal{M}^{theor}, w_i \models_t x : \alpha$ or $\mathcal{M}^{theor}, w_i \models_t x : \beta$;*
3. *$\mathcal{M}^{theor}, w_i \models_t \langle x, y\rangle : (\alpha \times \beta)$ iff $\mathcal{M}^{theor}, w_i \models_t x : \alpha$ and $\mathcal{M}^{theor}, w_i \models_t y : \beta$;*
4. *$\mathcal{M}^{theor} \models_t x : \alpha_a$ iff*
   - *$\mid W \mid \in \mathcal{M}^{theor} = n$;*
   - *$b = \mid w_i \mid \in \mathcal{M}^{theor}$ s.t. $\mathcal{M}^{theor}, w_i \models_t x : \alpha$;*
   - *$a = \frac{b}{n}$;*
5. *$\mathcal{M}^{theor} \models_t x : (\alpha \to \perp)_a$ iff $\mathcal{M}^{theor} \models_t x : \alpha_{1-a}$;*
6. *$\mathcal{M}^{theor} \models_t x : (\alpha + \beta)_a$ iff $\mathcal{M}^{theor} \models_t x : \alpha_b$, $\mathcal{M}^{theor} \models_t x : \beta_c$, and $a = b + c$;*
7. *$\mathcal{M}^{theor} \models_t \langle x, y\rangle : (\alpha \times \beta)_a$ iff $\mathcal{M}^{theor} \models_t x : \alpha_b$, $\mathcal{M}^{theor} \models_t y : \beta_c$, and $a = b \times c$;*
8. *$\mathcal{M}^{theor} \models_t [x]y : (\alpha \to \beta)_e$ iff*
   - *$\mid W \mid \in \mathcal{M}^{theor} = n$;*
   - *$c = \mid w_i \mid \in \mathcal{M}^{theor}\ s.t.\ \mathcal{M}^{theor}, w_i \models_t x : \alpha$ and $a = \frac{c}{n}$;*
   - *$d = \mid w_i \mid \in \mathcal{M}^{theor}\ s.t.\ \mathcal{M}^{theor}, w_i \models_t x : \beta$ and $b = \frac{d}{n}$;*
   - *and $e = a \times b$.*

We say that a model satisfies a non-probabilistic statement ($\mathcal{M}^{theor} \models_t x : \alpha$) iff this statement is satisfied in every world of this model. A possible world w satisfies $\Gamma = \{x^1 : \alpha^1, ..., x^n : \alpha^n\}$, denoted by $\mathcal{M}^{theor}, w \models_t \Gamma$ iff $\mathcal{M}^{theor}, w \models_t x^i : \alpha^i$ for all $i \in \{1, ..., n\}$. A model satisfies Γ, denoted by $\mathcal{M}^{theor} \models_t \Gamma$ iff the non-probabilistic subset of Γ is satisfied in every world w of this model, i.e., $\mathcal{M}^{theor}, w \models_t \Gamma$ for all $w \in W$, and the model satisfies all the probabilistic formulas occurring in Γ. A statement $x : \alpha$ is a semantical consequence of Γ, denoted by $\Gamma \models_t x : \alpha$, if $\mathcal{M}^{theor} \models_t \Gamma$ implies $\mathcal{M}^{theor} \models_t x : \alpha$. A statement $x : \alpha_a$ is a semantical consequence of Γ, denoted by $\Gamma \models_t x : \alpha_a$, if $\mathcal{M}^{theor} \models_t \Gamma$ implies $\mathcal{M}^{theor} \models_t x : \alpha_a$.

Empirical models

Definition 2. *Let $\mathcal{M}^{emp}$ be (W, R, v), such that W is non-empty set of worlds $w_1, ..., w_n$, $R \subseteq W \times W$ is temporal accessibility relation, $v : ES^t \to P(W)$ is a valuation function.*

1. *$\mathcal{M}^{emp}, w_i \models_e t : \alpha$ iff $w \in v(t : \alpha)$;*
2. *$\mathcal{M}^{emp}, w_i \models_e t : (\alpha + \beta)$ iff $\mathcal{M}^{emp}, w_i \models_e t : \alpha$ or $\mathcal{M}^{emp}, w_i \models_e t : \beta$;*
3. *$\mathcal{M}^{emp}, w_i \models_e t : \alpha \to \perp$ iff $\mathcal{M}^{emp}, w_i \not\models_e t : \alpha$;*
4. *$\mathcal{M}^{emp}, w_i \models_e \langle t, u\rangle : (\alpha \times \beta)$ iff $\mathcal{M}^{emp}, w_i \models_e t : \alpha$ and $\mathcal{M}^{emp}, w_i \models_e u : \beta$;*
5. *$\mathcal{M}^{emp} \models_e t_n : \alpha_f$ iff*
   - *$\mid W \mid \in \mathcal{M}^{emp} = n$;*
   - *$f = \mid w_i \mid \in \mathcal{M}^{emp}$ s.t. $\mathcal{M}^{emp}, w_i \models_e t : \alpha$.*

We say that a model satisfies a statement ($\mathcal{M}^{emp} \models_e t : \alpha$) iff the statement is satisfied in every world of this model. A possible world w satisfies $\Gamma = \{t^1 : \alpha^1, ..., t^n : \alpha^n\}$, denoted by $\mathcal{M}^{emp}, w \models_e \Gamma$ iff $\mathcal{M}^{emp}, w \models_e t^i : \alpha^i$ for all $i \in \{1, ..., n\}$. A model satisfies Γ, denoted by $\mathcal{M}^{emp} \models_e \Gamma$ iff the non-probabilistic subset of Γ is satisfied in every world w of this model, i.e., $\mathcal{M}^{emp}, w \models_e \Gamma$ for all $w \in W$, and the model satisfies all the probabilistic formulas occurring in Γ. A statement $t : \alpha$ is a semantical consequence of Γ, denoted by $\Gamma \models_e t : \alpha$, if $\mathcal{M}^{emp} \models_e \Gamma$ implies $\mathcal{M}^{emp} \models_e t : \alpha$. A statement $t_n : \alpha_f$ is a semantical consequence of Γ, denoted by $\Gamma \models_e t : \alpha$, if $\mathcal{M}^{emp} \models \Gamma$ implies $\mathcal{M}^{emp} \models_e t_n : \alpha_f$.

Fusion

Let $\mathcal{M}$ be defined as a couple $(\mathcal{M}^{emp}, \mathcal{M}^{theor})$.

Definition 3 (Satisfiability for $\mathcal{M}$)**.**

1. *$\mathcal{M}^{emp} \models_e \Gamma$ iff $\mathcal{M} \models \Gamma$, where Γ contains only expressions of $\mathcal{M}^{emp}$;*
2. *$\mathcal{M}^{theor} \models_t \Gamma$ iff $\mathcal{M} \models \Gamma$, where Γ contains only expressions of $\mathcal{M}^{theor}$;*
3. *$\mathcal{M} \models t_n : \alpha_{\tilde{a}}$ iff*
   - *$\mathcal{M}^{theor} \in \mathcal{M} \models_t x : \alpha_a$;*
   - *$\mathcal{M}^{emp} \in \mathcal{M} \models_e t_n : \alpha_f$;*
   - *$\tilde{a} = a \cdot n$;*
4. *$\mathcal{M} \models [x]t : (\alpha \to \beta)_{[a^*]b}$ iff for $|w| \in \mathcal{M}^* \subseteq \mathcal{M}$ s.t. $\mathcal{M}, w \models x : \alpha$ we have $\mathcal{M}^* \models y_t : \beta_b$.*
5. *$\mathcal{M} \models Trust(u_n : \alpha_f)$ iff $\mathcal{M}^{theor} \models_t x : \alpha_a$, $\mathcal{M}^{emp} \models_e u_n : \alpha_f$ and $a \in \epsilon(n)$, where $\epsilon(n)$ is the confidence interval for n;*
6. *$\mathcal{M} \models UTrust(u_n : \alpha_f)$ iff $\mathcal{M}^{theor} \models_t x : \alpha_a$, $\mathcal{M}^{emp} \models_e u_n : \alpha_f$ and $a \notin \epsilon(n)$, where $\epsilon(n)$ is the confidence interval for n.*

A statement Θ, where Θ is either $x : \alpha$, or $t : \alpha$, or $x : \alpha_a$, or $t_n : \alpha_f$, or $t_n : \alpha_{\tilde{a}}$ is a semantical consequence of Γ, denoted by $\Gamma \models \Theta$, if $\mathcal{M} \models \Gamma$ implies $\mathcal{M} \models \Theta$.

Examples

Imagine an AI process t which simulates throwing a die. An agent does not know whether this process is fair or not, but she knows that the theoretical probability of a fair die to get 3 is $\frac{1}{6}$. This ideal and fair die is represented by the model $\mathcal{M}^{theor}$ as on Figure 1. In this model, there exist six mutually exclusive and exhaustive worlds each of which contains one of six possible outcomes from a random variable associated with the process t.
In order to calculate the trustworthiness of t to produce 3, the agent executes a series of experiments. In particular, the process is activated 18 times, each time producing an output. The results of this experiment can be represented by the model $\mathcal{M}^{emp}$ as on Figure 2. In this model, t produces the output 3 exactly three times: during the 5th execution of t (world w_5), the 8th execution (world w_8), and the 18th execution (world w_{18}).

Figure 1.

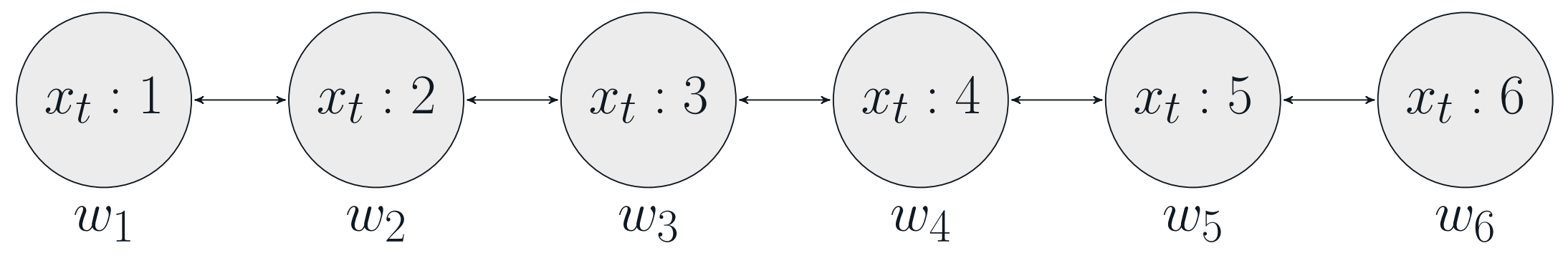
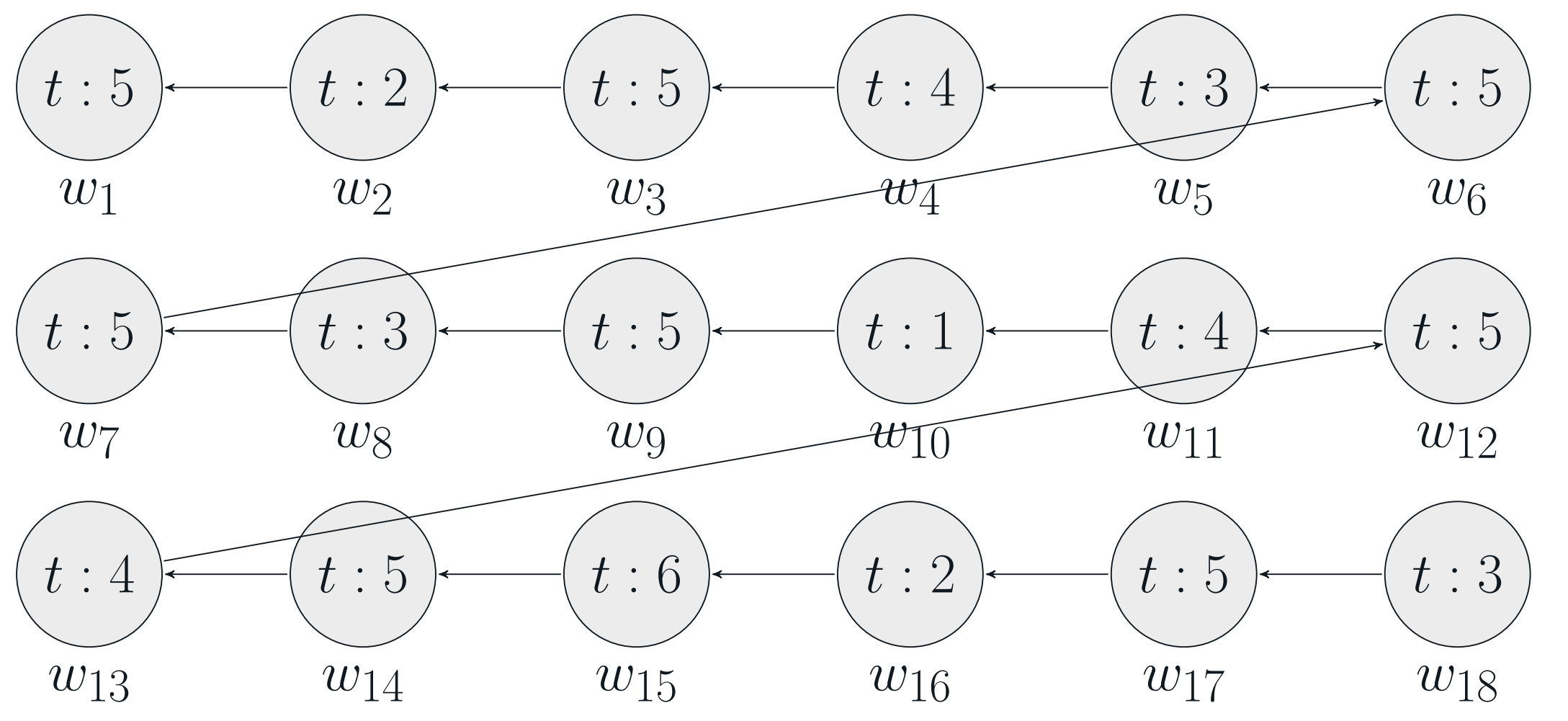


Figure 2.



Now the agent is able to evaluate the trustworthiness of t to produce 3. We have: $\mathcal{M}^{theor} \models x_t : 3_{\frac{1}{6}}$ and $\mathcal{M}^{emp} \models t_{18} : 3_3$.
Consider that the agent admits that, even if a process is trustworthy, the ideal probability of producing 3 and the real frequency of having 3 may not be the same. This is expressed by fixing the 95$\%$ confidence level for $\epsilon(18)$ under the normal approximation to the binomial distribution, which results in the interval $[-0.0055, 0.3389]$. Clearly, $\frac{1}{6} \in [-0.0055, 0.3389]$, and thus $\mathcal{M} \models Trust(t_{18} : 3_3)$.
Now let us consider the output 5. We have: $\mathcal{M}^{theor} \models x_t : 5_{\frac{1}{6}}$ and $\mathcal{M}^{emp} \models t_{18} : 5_8$.
In this case $\epsilon(18) = [0.2149, 0.6740]$. Having $\frac{1}{6} \notin [0.2149, 0.6740]$ we conclude $\mathcal{M} \models UTrust(t_{18} : 5_8)$.

References

- D'Asaro, F. A., Primiero, G. (2021). "Probabilistic typed natural deduction for trustworthy computations". In Dongxia Wang, Rino Falcone, and Jie Zhang, editors, *Proceedings of the 22nd International Workshop on Trust in Agent Societies* (TRUST 2021).
- D'Asaro, F. A., Primiero, G. (2022). "Checking trustworthiness of probabilistic computations in a typed natural deduction system". Authors' preprint.
- Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. (2018). *Ethics Guidelines for Trustworthy AI*.