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 Type 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.

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:
   - \( x \cdot \alpha \) means that a random variable \( x \) has value \( \alpha \);
   - \( x \cdot \alpha \) means that a process \( \Gamma \) produces an output \( \alpha \).

2. Expressions about the probability of a process to produce an output:
   - \( x \cdot \alpha \) means that the probability of a random variable \( x \) to have value \( \alpha \) is \( \alpha \);
   - \( x \cdot \alpha \) means that, after \( n \) executions of a process \( \Gamma \), an output \( \alpha \) was produced with frequency \( \alpha \);
   - \( x \cdot \alpha \) means that a process \( \Gamma \) produces an output \( \alpha \) with the expected probability \( \alpha \) over \( n \) executions.

TPTND contains four fragments:

- **Distribution construction rules**, which define the contexts as a list of assumptions on the probability distributions of processes to have certain outcomes;
- **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.

**Theoretical models**

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.

**Empirical models**

**Definition 2.** Let \( M^{\text{emp}} \) be \( (W, R, \ell) \), such that \( W \) is non-empty set of worlds \( w_1, \ldots, w_n \) such that \( w_i \cdot \alpha \) are mutually exclusive and exhaustive sets of \( ES^* \), \( R^{\text{emp}} \subseteq W \times W \) is an equivalence relation, and \( \ell : ES^* \to P(W) \) is a valuation function.

1. \( M^{\text{emp}} \models w_1 \cdot \alpha \land R \models \ell(w_1) \cdot \delta \rightarrow \epsilon \in (0; 1) \) if \( \epsilon \) is a valuation function.
2. \( M^{\text{emp}} \models w_1 \cdot \alpha \land \ell(w_1) \cdot \alpha \cdot \ell(w_1) : (\alpha + \beta) \models \ell(w_1) \cdot \alpha : \alpha \lor \ell(w_1) \cdot \beta : \beta \);\n3. \( M^{\text{emp}} \models w_1 \cdot \alpha \land \ell(w_1) \cdot \alpha \models \ell(w_1) : (\alpha + \beta) \models \ell(w_1) \cdot \alpha : \alpha \land \ell(w_1) \cdot \beta : \beta \);\n4. \( M^{\text{emp}} \models \ell(w_1) \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \);\n5. \( M^{\text{emp}} \models \ell(w_1) \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \).

We say that a model satisfies a statement \( M^{\text{emp}} \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \) if the statement is satisfied in every world of this model. A possible world \( w \) satisfies \( \Gamma = (t_1, \ldots, t_n) \), denoted by \( M^{\text{emp}} \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \).

Let \( M \) be defined as a couple \( (M^{\text{emp}}, M^{\text{theo}}) \).

**Definition 3 (Satisfiability for \( M \)).**

1. \( M^{\text{theo}} \models w_1 \cdot \alpha \models \Gamma \models M \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \);
2. \( M^{\text{theo}} \models w_1 \cdot \alpha \models \Gamma \models M \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \);
3. \( M \models \ell(w_1) \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \);
4. \( M \models \ell(w_1) \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \);
5. \( M \models \ell(w_1) \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \).

We say that a model satisfies a non-probabilistic statement \( M^{\text{theo}} \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \) if this statement is satisfied in every world of this model. A possible world \( w \) satisfies \( \Gamma = (t_1, \ldots, t_n) \), denoted by \( M^{\text{theo}} \models w_1 \cdot \alpha \models \ell(w_1) \cdot \alpha \land \ell(w_1) \cdot \beta : \beta \).

**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 \( t = \frac{1}{6} \). This ideal and fair die is represented by the model \( M^{\text{theo}} \) 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 \( n \) times, each time producing an output. The results of this experiment can be represented by the model \( M^{\text{emp}} \) 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 \( w_{18} \)).

**References**