Trustworthy AI — Jeannette Wing

June 30, 2025

The central questions we should be asking about AI are

- Why should we trust AI systems?
- How can we trust AI systems?

Formal verification of classical programs allows us to build (and be sure that we have) bug-free software. In this note, we discuss the differences between the verification of classical programs and AI, and mention several problems that arise in the formal verification of AI that are yet to be solved.

1 Impact of AI

AI has recently invaded all technological trends and is becoming (at least claimed) a priority of most big tech companies, governments, or research institutes.

1.1 AI capabilities

One of the reasons is that in various tasks AI can exceed human performance. Computers first beat the world chess champion in 1996 and since then, they improved a lot (see, e.g., [Sil+17]). When it comes to more practical topics, AI systems can quickly detect abnormalities in medical imaging, they are able to drive cars, or translate in real time.

1.2 AI can benefit humanity and society

AI has a lot of promise. We once again consider the medical imaging example. It would be greatly beneficial for our systems to be able to diagnose diseases quickly so that medical professionals can

help their patients more quickly.

1.3 Why should we trust AI systems?

Unfortunately, AI is quite flawed. Sometimes classifiers misclassify and hallucinate solutions. At the same time, AI is getting used in critical systems such as in medical facilities, avionics, and so on. So, this begs the question: how can we achieve trustworthy AI?

How can we achieve Trustworthy AI?

2 Trustworthy Computing

When it comes to a trustworthy system, there are several criteria that we can apply. System is trustworthy if it is the following list of properties (inspired by the definition of Microsoft):

- reliability (does it do the right thing?)
- safety (does it do any harm?)
- security (how is it vulnerable?)
- privacy (does it protect person's identity and data?)
- availability (is the system up when I need it?)
- usability (can humans use it easily?)

The preceding definition might be enough for classical systems. With AI, which is a probabilistic system by its nature, we have more demands (questions to answer):

- accuracy (how well does the AI systems do on new data?)
- robustness (sensitivity to a small change in the input)
- fairness
- accountability (who is responsible for the behavior of AI?)
- transparency (how does AI decide?)
- ethics (were the data used for training collected in an ethical manner?)
- ...properties yet to be discovered.

How can we achieve trustworthy AI? Through formal methods!

3 Probabilistic reasoning

Traditionally, we think of formal verification in terms of model checking. We ask whether

$$E, M \models P,$$

where M is a program code or an abstract model. $\models$ represents the logics and tools for checking the model such as model checkers, theorem provers, or SMT solvers. Lastly, P represents a formula (in a discrete logic) that represents wanted properties. E represents the system environment.

Model Checking AI
Thinking about AI, we have to "up the ante", and change our way of thinking about it. We change our model to the following:

$$D, M \models P$$

- D is our model of data, which can be a stochastic process or distribution that generates data inputs on which M's outputs need to be verified.
- M is our machine learning model.
- $\models$ could be interval analysis or probabilistic logics.
- P is formula expressing the desired property (this time in a logic that is not purely discrete, e.g., stochastic, probabilistic or discrete).

M is semantically and structurally different from a typical computer program:

- M is inherently probabilistic
- M is machine generated and human-unreadable ("intermediate code")
- it is a function of real numbers rather than over a discrete domain

P may be formulated over continuous domains

- Robustness properties for deep neural networks for deep neural networks are characterized as predicates over continuous variables.
- Fairness properties are characterized in terms of expectations with respect to a loss function over reals
- Differential privacy is defined in terms of a difference in probabilities with respect to a real value

$\models$ speaks about probabilistic logics and hybrid logics (these are logics mixing continuous and discrete case)

Conclusion: We need scalable or new verification techniques that work over reals, non-linear functions, probability distributions, stochastic processes, and so on.

Existing approaches:

- Hybrid automata – can reason about both discrete and continuous variables (see [Hen96])
- Differential dynamic logic (see [Pla08])

4 The role of data

available data = data at hand used to train
unseen data = data at which the model operates without seeing them in advance

How do we specify the unseen data?

- specify D as a stochastic process or data distribution or
- probabilistic programming language
- to specify unseen data, we need to make certain assumptions about the unseen data
- how is the unseen data related to the data used for training

4.1 The verification task $\models$

- How do we check the available data for desired properties? For example, if we want to detect whether a dataset is fair or not, what should we be checking about the dataset?
- If we detect that the property does not hold, how do we fix the model, amend the property, or decide what new data to collect for retraining the model?
- How do we exploit the explicit specification of unseen data to aid in the verification task?
- How can we extend standard verification techniques to operate over data distributions, perhaps taking advantage of the ways in which we formally specify unseen data?

5 Opportunities for Formal Methods

- There are opportunities for studying individual desired properties of AI.
- There is a need for devising compositional properties, allowing researchers to break down AI into components, speak about those, and then argue that the properties of the components hold for the whole.

6 Ethics

Belmont principles:

- Respect for persons (e.g. people should always be informed when they are talking to a chatbot)
- beneficence (risk/benefit analysis on the decision a self-driving car makes on whom not to harm)
- Justice (e.g., fairness of risk assessment tools in the court system and automated decision systems)

References

[Hen96] T.A. Henzinger. "The theory of hybrid automata". In: *Proceedings 11th Annual IEEE Symposium on Logic in Computer Science*. 1996, pp. 278–292. DOI: 10.1109/LICS.1996.561342.

[Pla08] André Platzer. "Differential Dynamic Logic for Hybrid Systems". en. In: *Journal of Automated Reasoning* 41.2 (Aug. 2008), pp. 143–189. ISSN: 1573-0670. DOI: 10.1007/s10817-008-9103-8. URL: https://doi.org/10.1007/s10817-008-9103-8 (visited on 07/06/2025).

[Sil+17] David Silver et al. *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. 2017. arXiv: 1712.01815 [cs.AI]. URL: https://arxiv.org/abs/1712.01815.