

Many recent research efforts have proposed innovative ML-based solutions for various cybersecurity problems. However, without understanding how these black-box models are making their decisions, cybersecurity researchers and practitioners have remained reluctant to trust them and have hesitated to deploy them in their production networks. A user trusts an ML model if that user is comfortable relinquishing control to the model. One key reason today's ML models cannot be trusted is that they are known to be vulnerable to underspecification issues, defined here as the failure to specify a model in adequate detail. Not unique to the cybersecurity domain, this problem manifests itself in ML models that are not credible because they exhibit unexpectedly poor behavior when deployed in real-world settings. An example of an underspecification issue is shortcut learning, where a model learns a shortcut (unrelated to the problem's causal structure) to make its decision. This observation has prompted growing interest in developing interpretable ML solutions (*e.g.*, decision trees) that can "explain" to humans why (and how) a given black-box model makes certain decisions (and not some other decisions) and, importantly when the model does and does not work. However, synthesizing such explainable models that capture a given black-box model's decision-making process with high fidelity while also being practical (*i.e.*, small enough for humans to comprehend) is challenging.

This paper presents Trustee, a novel framework that takes an existing black-box learning model and training dataset as input and generates a high-fidelity, easy-to-interpret decision tree explanation, and associated trust report as output. With this output, users can determine whether a given ML model of interest suffers from the problem of underspecification and can, therefore, not be trusted. To illustrate, the authors use Trustee to examine more than half a dozen of frequently cited and fully reproducible ML models from the existing cybersecurity literature and show that they all suffer from common instances of model underspecification, *e.g.*, evidence of shortcut learning, presence of spurious correlations, and vulnerability to out-of-distribution samples. The problematic nature of these findings serves as a cautionary tale as far as the widespread use of existing ML methods in the field of cybersecurity is concerned, argues for looking at developments in this area with a critical eye, and identifies specific pitfalls or "blind spots" that prevent users from trusting and deploying existing ML models in their production networks.

To encourage and enable other cybersecurity researchers and practitioners to examine and scrutinize proposed ML models to the point where they can make informed decisions on whether or not they can trust a given ML model, the authors built Trustee as a Python package that is compatible with most of the popular ML libraries and is publicly available at <https://pypi.org/project/trustee/>.

The paper received a Best Paper Honorable Mention at the 2022 ACM Conference on Computer and Communications Security (CCS'22) and won a 2023 Applied Networking Research Prize Award from IETF's Internet Research Task Force (IRTF).

## Basic Details

Arthur S. Jacobs, Roman Beltiukov, Walter Willinger, Ronaldo A. Ferreira, Arpit Gupta, and Lisandro Z. Granville. "AI/ML for Network Security: The Emperor has no Clothes." In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications

Security (CCS 2022). Association for Computing Machinery, New York, NY, USA, 1537–1551. <https://doi.org/10.1145/3548606.3560609>

Its reproducibility artifacts (e.g., source code, datasets, use cases) are publicly available on GitHub at <https://github.com/TrusteeML>. An extended version of the paper that includes further technical details and additional use cases is available at <https://github.com/TrusteeML/emperor/blob/main/docs/tech-report.pdf>.

The tool developed as part of this work (i.e., Trustee) is publicly available as a Python package at <https://pypi.org/project/trustee/>. It is compatible with the most popular ML libraries (e.g., PyTorch, Keras, Scikit-learn, and Tensorflow). Since its first release a few months ago, this package has already had over [6000](#) downloads.

## Recognitions

- This paper received a “**Best Paper Honorable Mention**” at ACM SIGSAC CCS 2022 and was the only paper from the “ML+Security” track to receive this recognition.
- This paper also received the 2023 “**Applied Networking Research Award**” from the Internet Engineering/Research Task Force (IETF/IRTF).