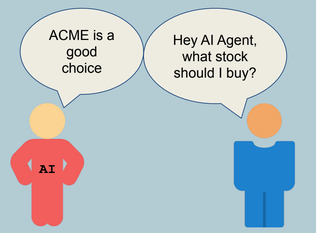
Verifiable AI

CENTER FOR ALGORITHMS AND INTERACTIVE SCIENTIFIC SOFTWARE

Paul Cesaretti
Rosario Gennaro
Tushar Jois
Hasan Kayman
Tugce Ozdemir

The City College of New York
The Grove School of Engineering

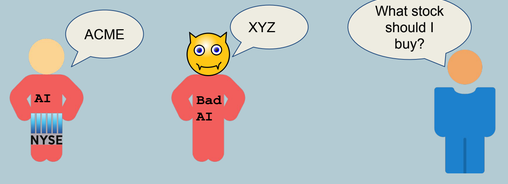
The Problem

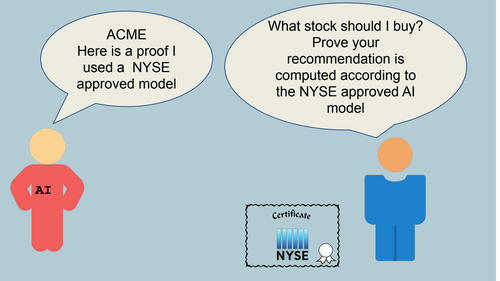


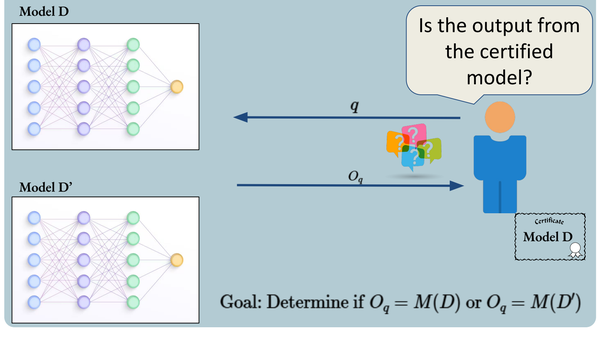
How can we ***trust*** the AI's recommendation?

We can assume the model has been certified by a "trusted authority"; But we do not run the model. Usually we query an Internet server (e.g. Google) who is supposed to run the right model.

How do we know if the certified model is being run?





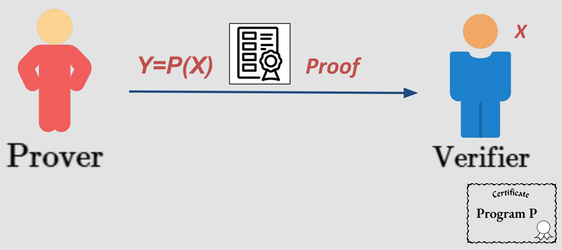


Goal: Determine if $O_q = M(D)$ or $O_q = M(D')$

Existing Technical Solutions

Cryptographic Proofs

Classical Result in Cryptography: Given any program ***P*** and an input ***X*** it is possible to prove that ***Y=P(X)*** to a "Verifier" who runs in much less time than re-executing ***P*** on ***X***



Technology is being used to verify the integrity of computations performed on ***untrusted cloud servers*** and to scale ***blockchain applications***

polygon
STARKWARE
RISC ZERO
Succinct
CLOUDFLARE
andreessen horowitz

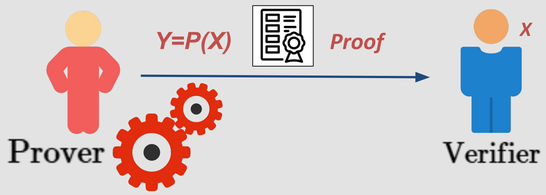
This includes startups working on the verification of the behavior of AI agents via the use of cryptographic proofs

Polyhedra
ModulusLabs

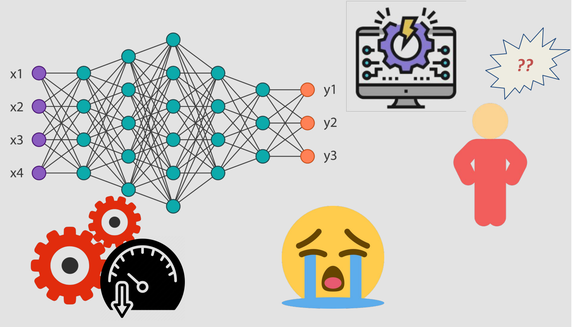
The Technical Problem

AI Models are too big for cryptographic proofs

To compute a proof the Prover must compute many ***expensive cryptographic operations***, a number that ***grows with the size of the program*** being proven

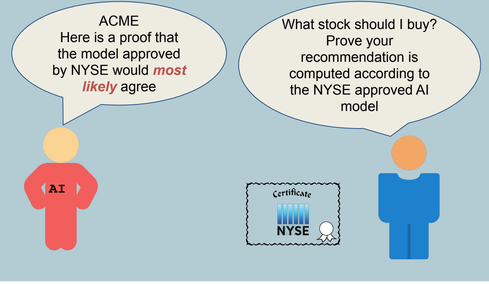


Neural Networks can be very large. High complexity: many neurons/connections. ***Too many cryptographic computations*** needed to produce a proof.



Our idea

Prove that the AI model is ***close*** to the certified one

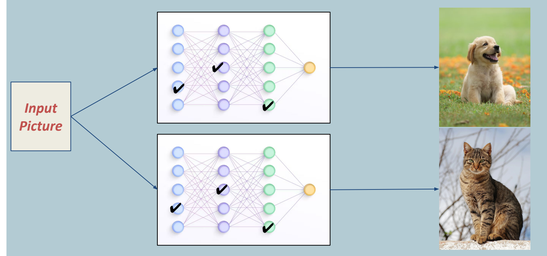


Our approach

Use statistical properties of neural networks

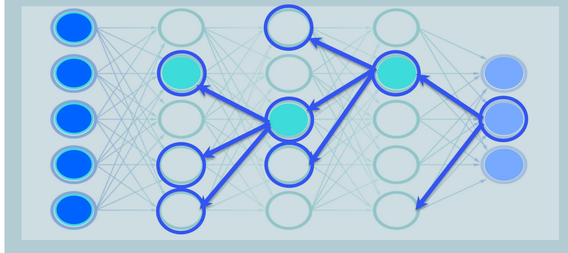
If two networks behave very differently, their neurons must follow very different activation distributions

Functional vs Representational Similarity of Neural Networks



Only checkmarked neurons have identical values, the others have different activations

Our conjecture: ***Random Spotchecking*** of internal neurons.
Instead of checking that all the neurons are activated according to the certified model, check the property for a random subset of appropriate size/structure.
This test should detect a prediction computed according to a very different model.



A cryptographic proof now must be run on a much ***smaller*** program, a randomly selected "sub-model", requiring ***a lot less expensive*** cryptographic operations

Acknowledgements

Research Funded by the Google Cyber NYC Institutional Research Program