Fairness in the Eyes of the Data

Certifying Machine-Learning Models

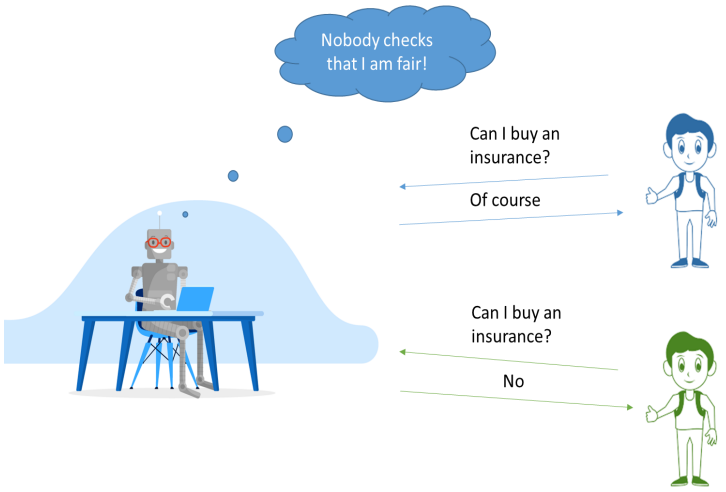
TEL AVIV UNIVERSITY

Shahar Segal Chaya Ganesh

Carsten Baum

Benny Pinkas
Yossi Keshet
Yossi Adi

Bar-Ilan University



Our setting [SAP+20]

The Client: has sensitive input ***x***. It wants to receive classification based on **a *fair model***.

The Server wants to **keep model *M* secret**. It might at be malicious and **could try to use an unfair model**.

The Server cannot just send *M* to the Client

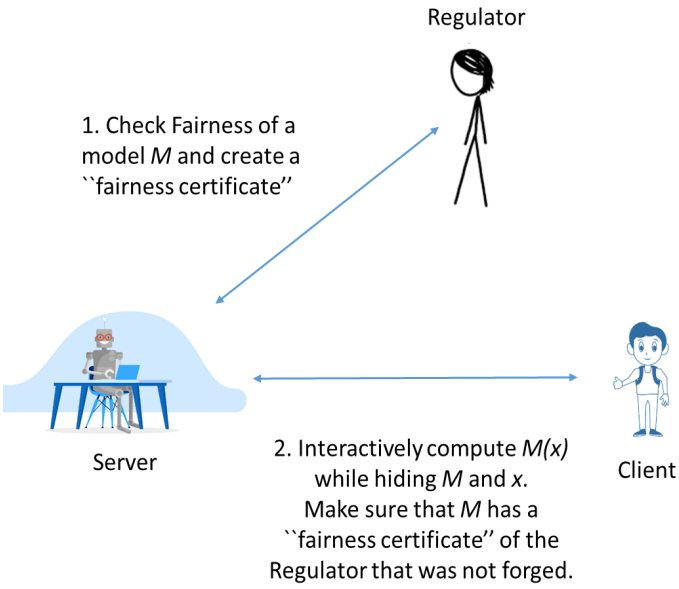
Introducing the Regulator

A fairness test of a model *M* requires special data and could be computationally expensive.

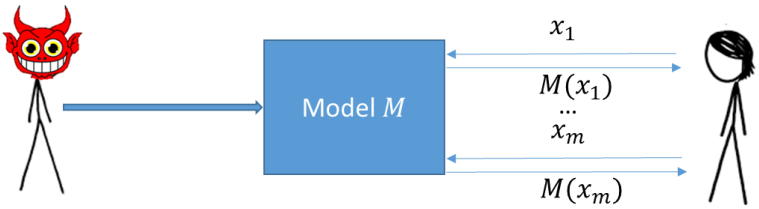
The Regulator certify ***fairness***. Certification only applies to checked models. A similar approach was taken in [KGK+18].

Protocol idea:

1. Test black-box *M* using Secure Computation. Regulator signs *M* if it is fair.
2. The Server and Client use Secure Computation during inference and check that *M* was certified.

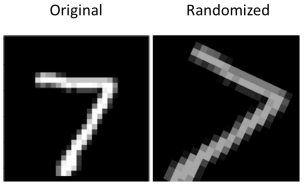


Testing for Fairness Black-Box



In our fairness test, the Regulator asks the Server to classify inputs. Nothing about *M* gets known to the Regulator.

We propose and empirically validate **a fairness test where the Server might even know the test data** before it creates *M*.



This test works by randomizing inputs such that their fairness-related attribute does not change under randomization.

References

[SAP+20] *Fairness in the Eyes of the Data: Certifying Machine-Learning Models*, Shahar Segal et al., https://arxiv.org/abs/2009.01534

[KGK+18] *Blind Justice: Fairness with Encrypted Sensitive Attributes*, Niki Kilbertus et al., ICML 2018