TRUST4AI.FAIRNESS

Propose fairness measures for cybersecurity products based on AI solutions. Fairness works: Detect and mitigate bias in AI models.

AI4CYBER

Component

The goals is to ensure that cybersecurity systems do not introduce bias and discrimination in the process on AI analyses. The proposed works are a good differentiator for a product or service offer: very few studies exist on the subject and works ensure compliance with EU AI Act.

The AI4CYBER framework is being designed to leverage Artificial Intelligence (AI) capabilities to enhance cyber resilience of critical systems. The framework offers a set of new generation AI-based services that support cyber security and robustness of critical systems in a more efficient way.

The TRUST4AI component in AI4CYBER framework is composed of three sub-components, namely: the subcomponent TRUST4AI.XAI is dedicated to the interpretability of ML and AI systems, the subcomponent TRUST4AI.Fairness is about Fairness of ML and AI systems while TRUST4AI.Security enforce Security of others AI4CYBER's components. TRUST4AI.Fairness provides the others AI4CYBER components tools dedicated to fairness in ML and AI. The fairness may be described by considering two distinct notions: disparate treatment and disparate impact.

TRUST4AI.Fairness is based on three components to detect and mitigate bias.

- Fairness metrics: The component proposes several types of metrics (base rate metrics, group accuracy and calibration metrics, individual fairness). Each family of metrics aims to measure something different (bias in the models' outputs, bias in model errors, etc.).
- Bias mitigation: Bias can be mitigated at three levels: pre-processing, where the biases are mitigated before the training of ML model, in-processing where the bias correction is integrated directly in the training process and post-processing where the biases are corrected during model prediction. For TRUST4AI.Fairness, we will focus on the pre and post processing ones as the tools implemented should be used by the others AI4CYBER components with as little as possible change.
- Dashboard: The AI Fairness Dashboard is a comprehensive tool designed to measure and mitigate bias in AI models.

YouTube Video Link

License – open source // commercial license
Thales Internal.

THALES

Cong-Bang Huynh (cong-bang.huynh@thalesgroup.com)
Stéphane Lorin (Stephane.lorin@thalesgroup.com)
Vincent Thouvenot
(Vincent.thouvenot@thalesgroup.com)
Thales Gorup

https://www.thalesgroup.com/en

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450. Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.