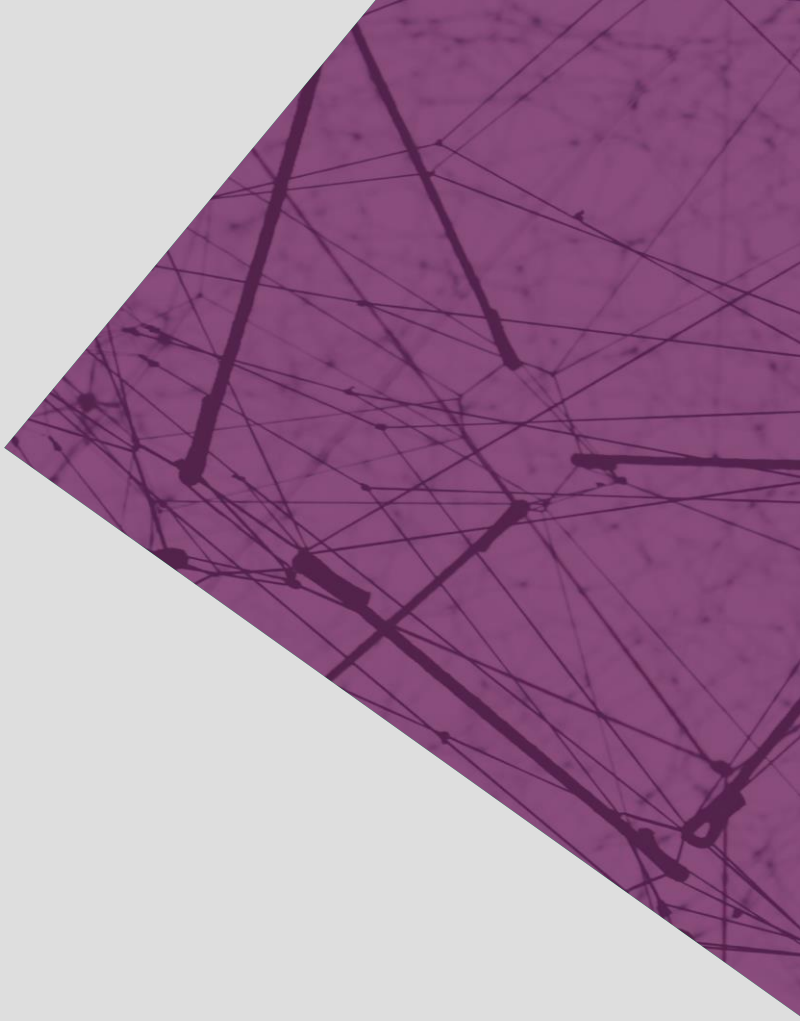
The Alan Turing Institute



Robustness

Milestone 5: Trade-offs and Interactions with other verticals in Trustworthy AI

Roseline Polle

Postgraduate, UCL

roseline.polle.19@ucl.ac.uk

What is Robustness?

Technical Robustness and Safety (EU guidelines):

- Resilience to attack and security
- Fallback plan and general safety
- Accuracy
- Reliability and Reproducibility

EU-HLEG. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Resilience to attack and security

- Quality of a system to be safe, not vulnerable to tampering.
- Protect against hacking: data poisoning, model leakage or the infrastructure, both software and hardware.

Type of ML attacks

- Integrity: Misclassifications that do not compromise normal system operation (evasion, poisoning,…)
- Availability: Misclassifications that compromise normal system operation (poisoning)
- Privacy/Confidentiality: infer information about user data and models.

Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning."

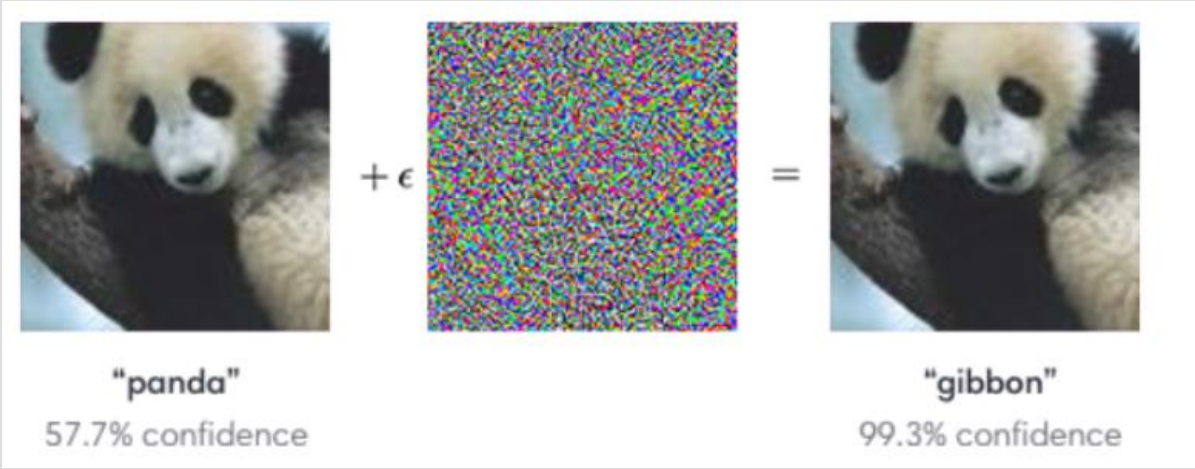
Example of ML attacks

- *Evasion attacks*: manipulating input data to evade a trained classifier at test time

- *Poisoning attacks:* injecting a small fraction of poisoning samples into the training data (occur during the training phase) to increase misclassification at test time.

Small changes in input

Small changes in input should lead to small changes in output



[Goodfellow et al., 2015]

Fallback plan and general safety

- Safeguards that enable a fallback plan in case of problem
- Continue operation with minimisation of unintended consequences and errors: human-in-the-loop, switching to rule-based,…

Accuracy

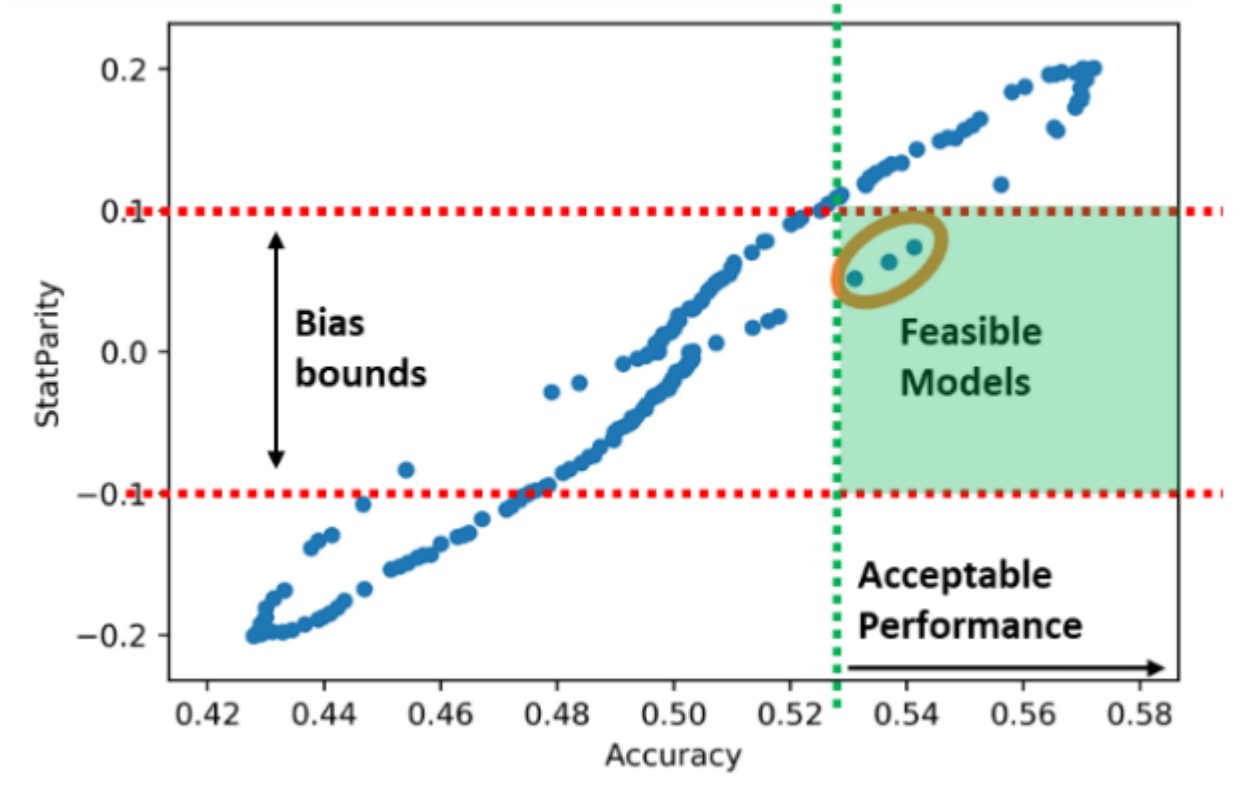
- High level of accuracy desirable
- Explicit and well-formed development and evaluation process

Reliability and Reproducibility

- Works properly with a range of inputs and in a range of situations
- Same behaviour when repeated under the same conditions

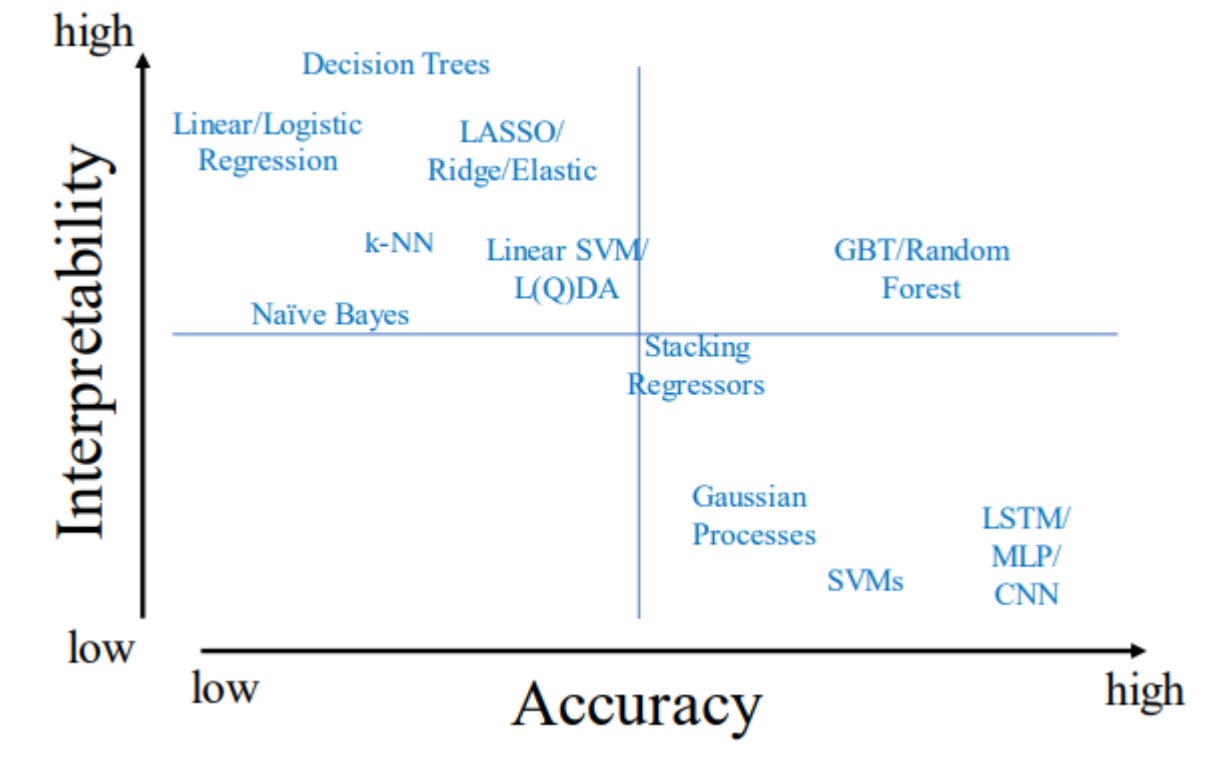
Robustness and Fairness

Bias vs Accuracy



Towards Algorithm Auditing by Koshiyama et al., 2021

Accuracy vs Explainability



Towards Algorithm Auditing by Koshiyama et al., 2021

Further readings

- *Why Robustness is not Enough for Safety and Security in Machine Learning* by Christian Kästner
- Goodfellow et al., *Explaining and Harnessing Adversarial Examples*