7 Trusted AI & Ethics

Trusting AI systems

- Any practical AI system in production needs to be:

- Fair
  - Not allowing for any bias or discrimination
- Robust
  - Not able to be manipulated from the outside
- Explainable
  - Able to understand the internal decision process

- Need for AI governance and responsible AI
  - Technical solutions exist, but at some costs (e.g., slower execution)

Fairness

- No discrimination against minorities or bias in decisions
- Bias is often present in data and transferred into models
  - Toxic effects of reinforcing existing unhealthy stereotypes
- Some recent examples
  - Facial recognition worked better for light-skinned males (Buolamwini)
  - Man is to computer programmer as women is to homemaker? (Bolukbasi)
  - Amazon’s hiring tool discriminated against women (Reuters)

Robustness

- Not able to be manipulated by a third party via adversarial attacks
  - Deliberately force to make a wrong prediction and trying to fool the AI

- Make the system do something else than it is intended to do:
  - Stickers on stop sign confuse the AI
  - Patch that tricks AI into thinking a banana is a toaster
  - Glasses make facial recognition AI think you’re actress Milla Jovovich

- Adversarial use of AI
  - Obama Deep Fake video

Explainability

- Understand why a specific decision is made
  - User has the “right to an explanation” (GDPR)
  - Especially important for high-stakes decisions with a big impact on lives
- Wolf vs husky experiment (Ribeiro et al.)
  - Snow in the background? → Husky
- Two options to guarantee explainability
  - Tranparent models
  - Ex-post interpretation techniques of black box models (many exist)