kokchun giang

approaching AI with **ethics** to build a fair and smarter world

AI systems learns from human data and **scales prejudices** in unprecedented speed

COMPAS recidivism (2016)

predict likelihood of reoffending

black defendants more incorrectly labeled as high risk

white defendants more incorrectly labeled as low risk

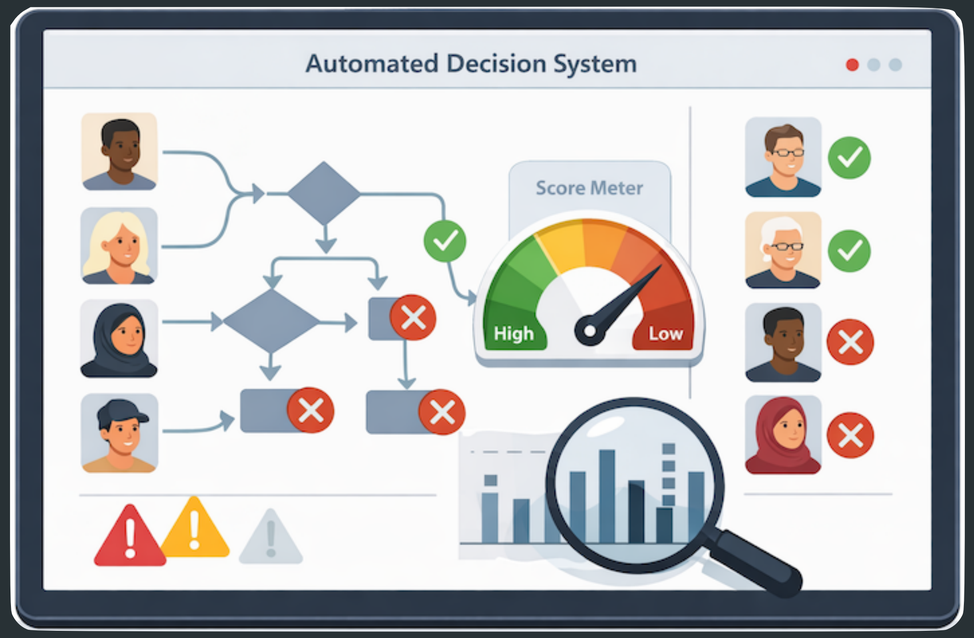
Amazon recruiting (2018)

AI tool to review resumes

trained on past 10 years submitted resumes

male dominated industry
→ AI learnt that male is success factor

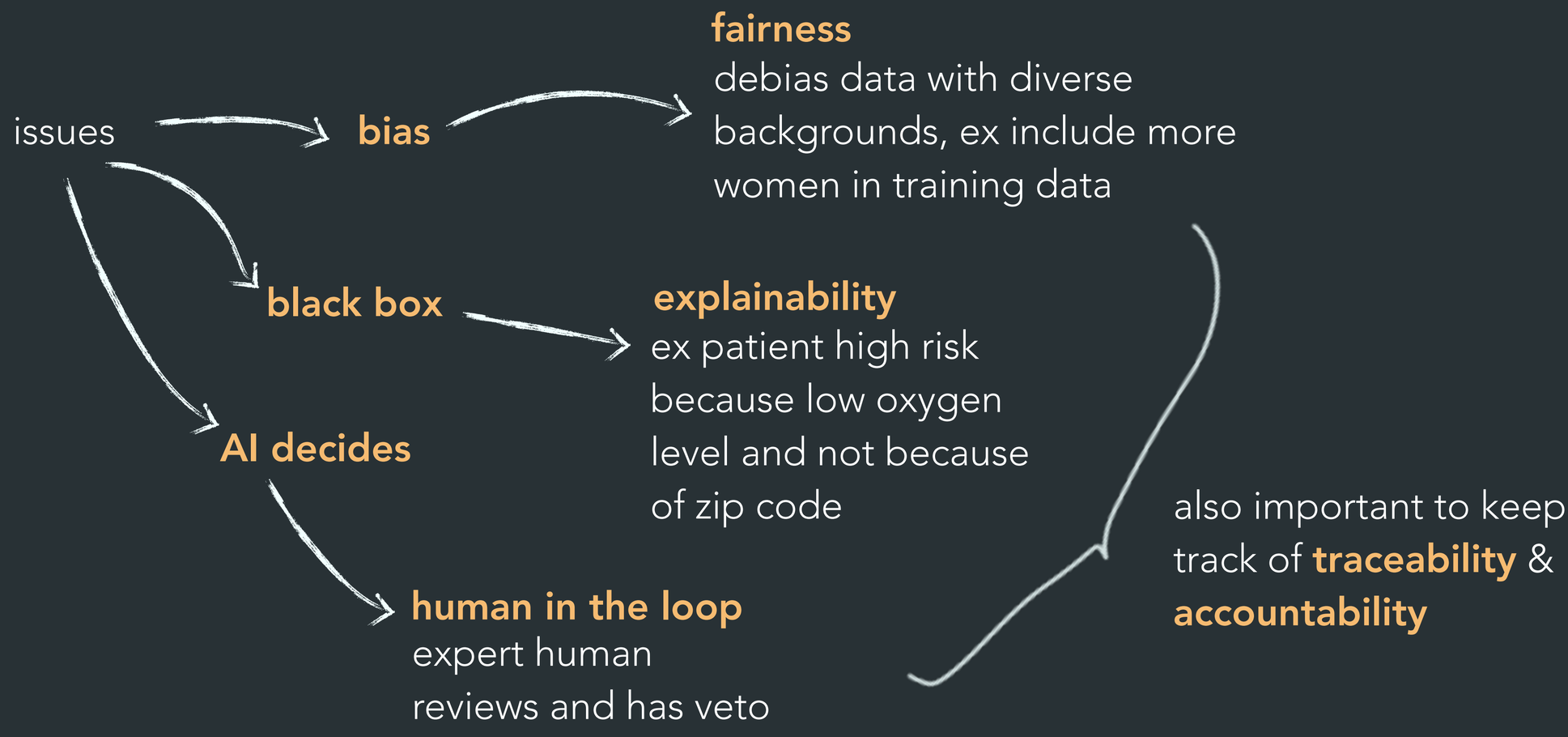
the **scale** is massive when an AI tool is biased



a human recruiter that is biased may affect hundreds of people

an AI recruiter that is biased may affect millions of people

AI ethics to mitigate some of the failures



intentionally using AI for **unethical purposes**

AI is a powerful tool, but can be used for unethical purposes

deepfakes to make someone say something not true

automating and scaling financial fraud and deception

mass surveillance

mass discrimination

aid in cybersecurity crimes

voice cloning scams

disinformation & election interference

systemic discrimination

intellectual property theft

policy landscape - innovation vs regulations?

GDPR
2018

data privacy

China algorithm regulation
2022-23

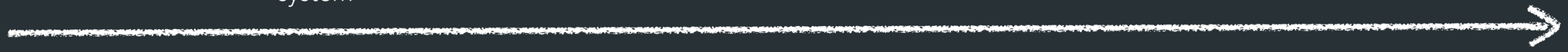
watermarks on deepfakes, user opt out from recommendation system

EU AI act
2024

comprehensive law, uses risk-based approach. Bans unacceptable AI

AU AI strategy
2024

African Union framework focused on sovereignty, avoiding African languages & cultural data to be exploited



UNESCO recommendation
2021

193 countries adopted, ban AI for social scoring

US exec order
2023

AI companies share safety results with US government

Global accord
23, 24

mitigate catastrophic risks, AI respects human rights

US exec order
2025

focus on innovation and reduce regulatory