MAKING AI DECISIONS TRANSPARENT

LET THERE BE LIGHT

BRINGING AI INTO THE LIGHT

Transparency in AI Decisions

XAI aims to make AI decisions transparent and understandable, enhancing trust in technology.

High-Stakes Applications

Clear AI reasoning is essential in fields like healthcare, law, defense, and finance to ensure reliable outcomes.

Accountability and Ethics

XAI fosters accountability and ensures ethical use of AI technology by making its decisions visible.

WHY AI IS HARD TO UNDERSTAND

Complex AI Models

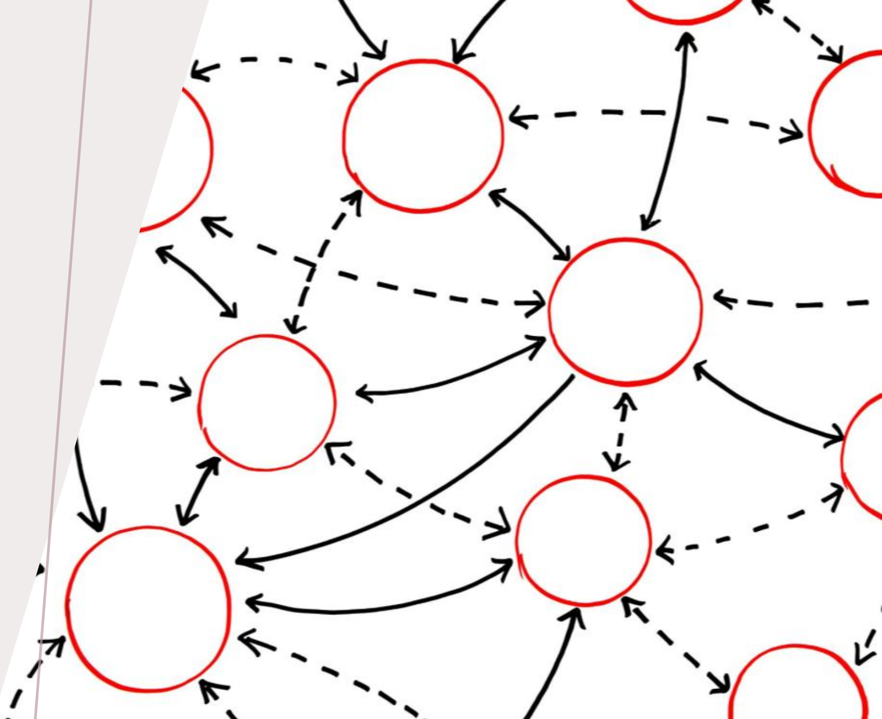
AI models contain billions of parameters, making them very complex and difficult to interpret, resembling a 'black box'.

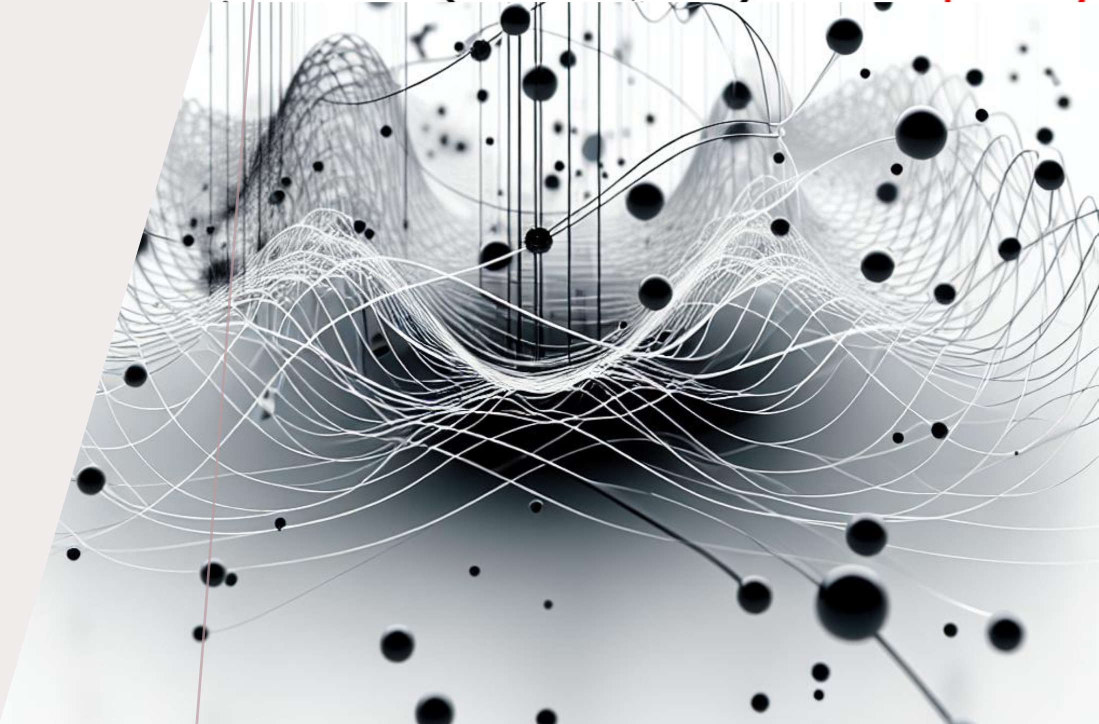
Non-linear and Emergent Outputs

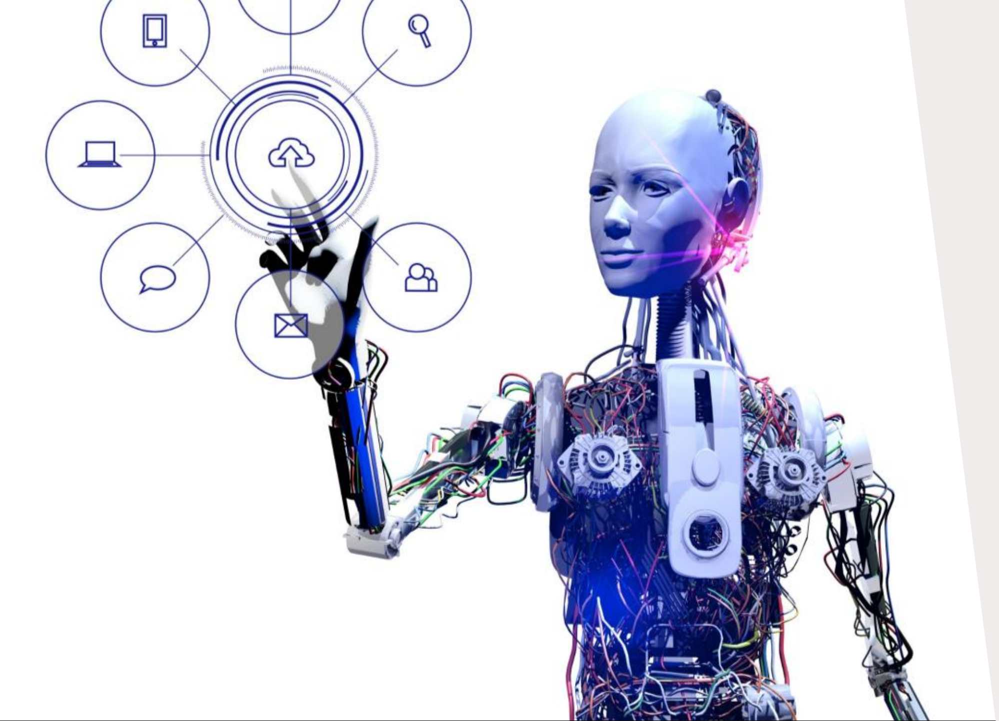
AI outputs are non-linear and emergent, leading to unpredictable behavior that is hard to anticipate.

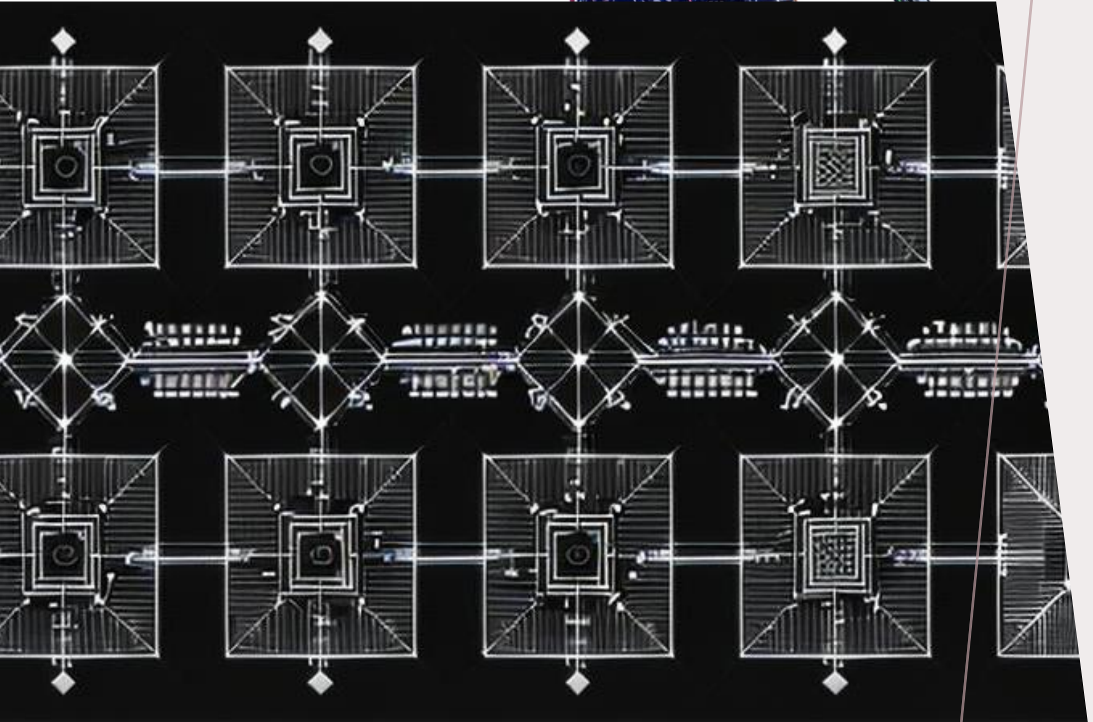
Post-hoc Explanations and Bias

Explanations for AI decisions are often made post-hoc, and biased training data can result in biased AI outcomes, affecting fairness and accuracy.









CAN WE TRUST THE MACHINE?

Performance Metrics

F1, Precision, and Recall are used to measure the accuracy and quality of machine learning models.

Adversarial Testing

Adversarial testing evaluates a model's robustness against malicious inputs to ensure reliability under attack.

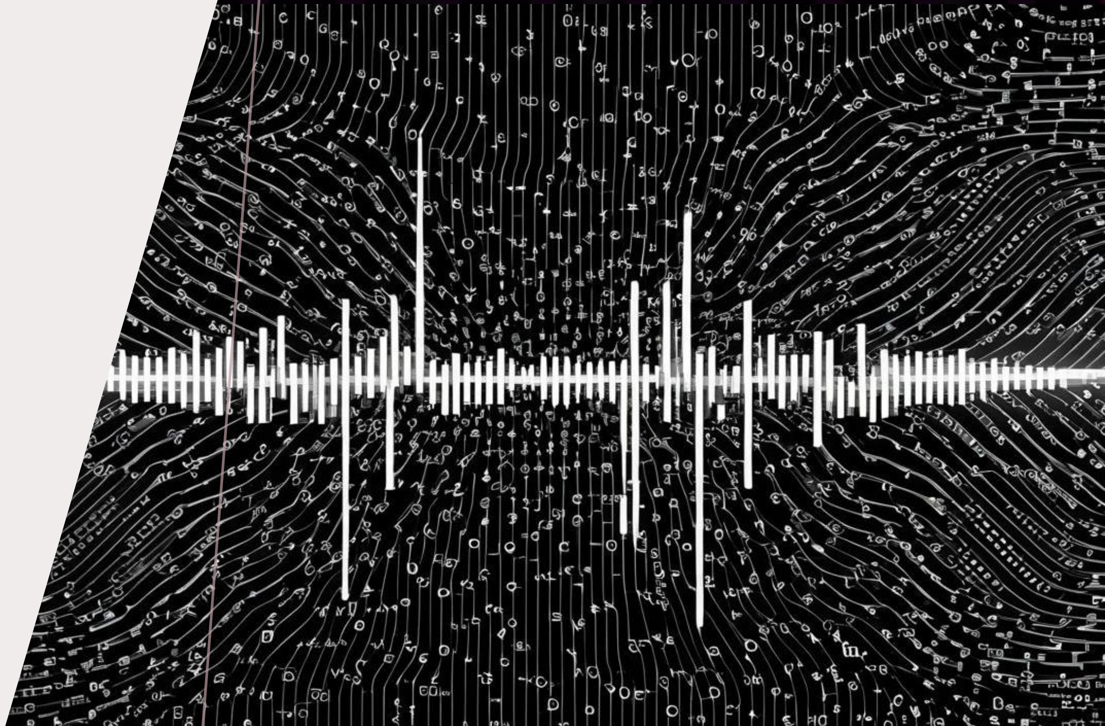
Human Evaluation

Human evaluation is conducted to ensure that the results produced by the model make practical sense.

Fairness Audits

Fairness audits check the model for biases to maintain social trust and ensure equitable outcomes.

LEADING AI LABS' SOLUTIONS



OpenAI's Evaluation Tools

OpenAI develops evaluation tools and gathers user feedback to continuously improve its AI models' performance.

Anthropic's Constitutional AI

Anthropic advances Constitutional AI, focusing on creating AI systems that adhere to ethical guidelines and principles.

Google's Transparency Enhancements

Google enhances AI transparency by providing visual trace explanations, making AI decisions more understandable.

Meta's Weight Analysis

Meta focuses on open weight analysis, promoting accountability and advancements in AI model development.

EXPLAINABILITY + METRICS = TRUSTWORTHY AI

AI Explainability

Explainability helps users understand AI decision-making, making systems more transparent and usable.

AI Model Validation

Validation ensures AI model reliability and accuracy through thorough testing and verification processes.

Performance Metrics

Metrics quantify AI performance, identify biases, and ensure consistent evaluation standards for accountability.