Faysal Aziz

AI Safety & Ethics

From Circuits to Secrets in a day

FA-0521

Agentic & Generative AI · DevOps, Cloud & Infrastructure

Introduction

Understand how AI works, why it's unpredictable, and how to keep it safe in a world that's playing catch-up.

AI is no longer a niche research field—it's a mainstream force shaping politics, industry, and everyday life. Yet the people building and regulating it often understand it only partially. This course blends an accessible technical primer with a candid look at geopolitical tensions, regulatory loopholes, and ethical dilemmas. Led by an instructor with over 30 years of industry experience, we'll combine serious insights with engaging, relatable examples—from AI-driven global rivalries to deepfake mischief—to explore how to use AI responsibly in the real world.

Learning outcomes

By the end of the day, participants will:

- Understand the basics of how AI systems—especially large language models (LLMs)—are trained and operate.
- Explain why AI behaves unpredictably and remains a "black box" despite impressive capabilities.
- Describe key global AI political and regulatory positions, including the EU's military exemption and the Trump administration's deregulatory stance.
- Recognize the role of U.S.–China AI rivalry in shaping technology access and standards.
- Evaluate the ethical and legal challenges of AI training data collection without consent.
- Identify threats from deepfakes and apply safety measures to avoid being fooled.
- Engage critically and creatively with AI ethics issues through interactive activities.

Prerequisites

- Basic familiarity with AI concepts and current events.
- No coding knowledge required—just curiosity and critical thinking.

Training outline

1. 1. AI 101: Under the Hood

What AI Really Is:

- Narrow AI vs. General AI.
- Neural networks, machine learning, and deep learning—how they differ.

How AI Learns:

- Data collection, preprocessing, and labeling.
- Training using backpropagation and gradient descent.
- Role of GPUs/TPUs in computation.

Large Language Models (LLMs):

- Tokenization, embeddings, attention mechanisms, and transformers.
- The scaling laws that made modern AI possible.
- Why bigger data + more compute ≠ always better results.

2. 2. AI's Mystery Sauce

The Black Box Problem:

- Why models can solve problems without us fully knowing *how*.
- Complex feature representations that humans can't intuitively interpret.

Geoffrey Hinton's Concerns:

- His suspicion of backpropagation's long-term viability.
- Calls for rethinking AI foundations before unexpected failures emerge.

Non-Determinism in Action:

- Why repeated prompts to the same model can produce different outputs.
- Impact on safety, reliability, and accountability.

Implications:

- Limits of interpretability tools like SHAP and LIME.
- Risk assessment in high-stakes environments.
- Overview of recent AI safety incidents: deceptive LLM behavior and evaluation shortcomings

3. 3. Geopolitics & Regulation

Key global initiatives:

- EU's AI Act (in force since August 1, 2024) and applicability timelines
- UNESCO's AI Ethics recommendations
- California's policy report on irreversible AI risks and trust-but-verify approach
- Other governance efforts (AI Safety Institutes, global summits)

- **EU's Double Standard**: Civilian AI oversight vs. military exemptions.
- **Trump's Deregulation Push**: Prioritizing AI dominance over safety rules.
- **U.S.–China Cold War**: Chips, data, and digital infrastructure supremacy.

4. 4. Data Ethics

- **The Consent Problem**: How most LLMs used public data without permission.
- **Legal Grey Zones**: GDPR's "legitimate interest" loopholes vs. privacy rights.
- **Cultural Impacts**: Artists, journalists, and domain experts losing control of their work.

5. 5. Deepfakes

- **How They Work**: GANs, diffusion models, and real-time manipulation.
- **Real Dangers**: Scams, political disinformation, revenge porn.
- **Personal Safety**: Verification habits, detection tools, skepticism training.
- **Fun Demo**: Light-hearted but safe deepfake examples to illustrate risk without harm.

6. 6. Building Ethical Assurance into AI Systems

- Frameworks for ethics-based auditing and continuous oversight.
- Implementing transparency, explainability, accountability, and governance layers aligned with best practices and standards.

  Discussion: evaluating tools like

  - HELM Safety,
  - AIR-Bench,
  - FACTS for factuality and safety testing

7. 7. Interactive Wrap-Up

- **Scenario Game**: Choose your role—regulator, engineer, activist, or everyday citizen—and craft a mini AI safety plan.

- **Meme-Making**: Teams create ethical memes from course takeaways.
- Group reflection on AI’s future and where responsibility truly lies.

© 2026 Faysal Aziz · Kuala Lumpur