Future of AI, AI Fairness & LLM Safety

IMPORTANT NOTES:
Study lecture materials at least 1 hour and prepare the questions prior to the tutorial session.
The questions will be discussed in the tutorial session.

1. From the fairness shapley value, if we observe that a certain feature contributes a lot to the unfairness (e.g., marital status in demographic parity difference), is it always correct to remove that particular feature from the model?

   No, removing the feature is not always correct. The feature may still provide predictive value, and correlated proxies can reintroduce unfairness. Instead, fairness should be addressed through domain-specific rules and mitigation methods (e.g., reweighting, fairness-aware algorithms) rather than simply dropping the variable.

2. Why is fairness a critical consideration in AI systems, especially in healthcare applications?

   Fairness ensures that AI models do not perpetuate or amplify existing discrimination in healthcare delivery. For instance, biased data may cause underdiagnosis in certain demographic groups. Without governance, AI risks worsening health inequities, making fairness central to safe deployment.

3. If an AI system in healthcare makes a biased diagnosis due to biased training data, who should be held responsible — the developers, the healthcare provider, or the regulator?

   Responsibility is shared:

   - Developers must ensure training data and models are robust against bias.
   - Healthcare providers must validate AI recommendations before adoption.
   - Regulators must enforce standards that protect patients.

   Ultimately, accountability should be layered, with developers ensuring technical safeguards and clinicians retaining final decision-making authority.

4. Do you think fairness in AI should be defined universally, or should it adapt to local cultural and social contexts?

   While some fairness principles (e.g., non-discrimination) should be universal, fairness often depends on cultural values. For instance, prioritizing equality of outcome vs. equality of opportunity may vary across societies. A hybrid approach works best: universal guardrails with local adaptations.

5. What makes AI governance especially challenging compared to traditional technology governance?

   Challenges include lack of an evidence base, fragmented stakeholder interests, low technical literacy among policymakers, and amplified global inequalities. Governance must balance innovation with harm reduction across diverse regulatory landscapes.

6. With AI systems capable of creating realistic deepfakes, should governments regulate the use of generative AI or the development of the underlying models?

   Regulating use is more practical, as banning development may stifle innovation. However, safeguards at the model level (e.g., watermarking, provenance checks) should be mandatory. A dual approach — controlling misuse while embedding safety-by-design — balances innovation and protection.

7. Imagine a future where AI systems become indistinguishable from humans in voice, face, and behavior. How might this reshape human trust, law, and governance?

   Trust systems would need to shift from perception-based ("I saw it, so it's true") to verification-based (cryptographic proofs, digital watermarks). Legal frameworks would need to define AI personhood boundaries and establish liability for harm caused by AI-generated actions. Governance would likely mandate identity verification in critical systems.

8. What are the main risks associated with deploying LLMs in real-world applications?

   Risks include:

   - Hallucination (LLMs generating plausible but false information).
   - Bias propagation (reproducing or amplifying societal stereotypes).
   - Prompt injection (adversarial inputs manipulating outputs).
   - Privacy leakage (unintended disclosure of sensitive training data).
   - Over-reliance by users (blind trust in model outputs).

9. How can we prevent LLMs from leaking sensitive information embedded in their training data?

   - Data filtering during pre-training to exclude sensitive data.
   - Differential privacy to prevent memorization of personal data.
   - Red-teaming & safety audits to test for data leakage.
   - Access control & use policies limiting sensitive domains.

10. LLMs often struggle with adversarial attacks such as prompt injection. Why is this problem particularly difficult to solve?

    LLMs are trained to follow instructions and continue text; adversarial prompts exploit this very design. The open-ended nature of natural language makes it hard to define "safe" vs. "unsafe" instructions in advance. Even with filtering and guardrails, attackers can obfuscate malicious requests (e.g., using metaphors or code words)

11. What role do "red-teaming" and adversarial testing play in LLM safety?

    They simulate malicious or unexpected use cases to uncover vulnerabilities. Provide feedback loops for improving safeguards before public release. Enable stress-testing across cultural, ethical, and linguistic contexts. Help establish benchmarks for responsible AI deployment.

12. What governance mechanisms can be put in place to ensure LLM safety at scale?

    - Transparency requirements (model cards, system documentation).
    - Independent auditing of model safety and fairness.
    - Regulatory alignment with frameworks like the EU AI Act.
    - Ethics review boards for high-risk deployments.
    - Global collaboration to prevent regulatory arbitrage.