Defense vs Generative AI Fakes and Bias

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. A cybersecurity team has noticed an increase in sophisticated phishing emails targeting their organization. These emails contain perfect grammar, personalized content, and current events references.

   (a) Explain how generative AI makes phishing attacks more dangerous compared to traditional phishing.

   (b) Compare the differences between AI-enabled automated phishing, spear phishing, and vishing attacks.

   (c) What defensive measures would you recommend to protect against AI-generated phishing campaigns?

   (a) How generative AI makes phishing more dangerous:

   - Improved linguistic quality: AI eliminates grammatical mistakes and produces professional writing styles that were traditionally red flags.
   - Hyper-personalization: LLMs can absorb real-time information from news outlets, corporate websites, and social media to create highly targeted content
   - Scale and speed: AI chatbots can create and spread business email compromise campaigns much faster than humans
   - Current context integration: Incorporation of up-to-the-moment details makes emails appear more legitimate

   (b) Comparison of AI-enabled attack types:

   - Automated Phishing: Mass targeting; Moderate personalization based on available data; Generic but sophisticated emails sent to many recipients
   - Spear Phishing: Specific individuals/organizations; Highly customized using extensive research; Appears from trusted sources like colleagues/friends
   - Vishing: Individual targets via voice; Voice cloning of trusted contacts; Phone calls/voicemails using deepfake audio

   (c) Defensive measures:

   - Implement multi-factor authentication beyond phishable methods
   - Use deterministic verification strategies instead of probabilistic ones
   - Regular security awareness training updated for AI-generated threats
   - Email filtering systems that detect AI-generated content patterns
   - Verification protocols for unusual requests, especially financial transfers
   - Voice authentication systems that can detect deepfake audio

2. A healthcare organization wants to share patient data for research while maintaining privacy. They're considering implementing differential privacy using the RAPPOR technique.

   (a) Explain the core principle of differential privacy and how it protects individual privacy.

   (b) Walk through the RAPPOR randomized response technique using a healthcare example.

   (c) What is the longitudinal privacy problem, and how does Permanent Randomized Response solve it?

(a) An algorithm is differentially private if outcome statistics don't reveal whether a specific individual was in the dataset. By adding calibrated random noise to original data, it achieve protection mechanism, such that individual users become unidentifiable while preserving statistical utility. Note that the noise averages out over large datasets, maintaining valuable insights while protecting individuals.

(b) RAPPOR technique - Healthcare scenario: Determining percentage of patients with a sensitive condition (e.g., mental health issues).
Process:

i. Setup: Survey 1,000 patients about depression diagnosis

ii. Randomization: Each patient privately rolls a dice

iii. Response rule:
   - If dice = 6 (probability p = 1/6): Answer Question 1 truthfully
   - If dice = 1-5: Answer Question 2 truthfully
   - Q1: "I have been diagnosed with depression"
   - Q2: "I have never been diagnosed with depression"

Calculation:

- Let T = true percentage with depression
- Let S = surveyed percentage saying "yes"
- Formula: $S = pT + (1 - p)(1 - T)$
- Solving for $T : T = (S + p - 1)/(2p - 1)$
- Example: If $S = 40\%$ and $p = 1/6$, then $T = (0.4 + 0.167 - 1)/(0.333 - 1) = 65\%$

(c) Longitudinal privacy problem and solution: Repeated one-time randomized responses on same person can reveal identity over time. Sequence of responses creates pattern that attackers can analyze
Permanent Randomized Response solution:

- Replace one-time randomization with permanent noisy value.
- Each individual gets assigned a consistent randomized response (e.g., '1', '0', or original response).
- Example: Patient reporting daily symptoms uses same noise pattern throughout study period.
- Benefit: Attackers cannot differentiate between true and noisy data in longitudinal datasets.

3. Bias Detection in Generative AI: You're auditing a generative AI system for bias issues. The system is used for hiring recommendations and content generation.

   (a) Identify and explain the three main types of biases that can occur in AI systems, providing examples for each.

   (b) How would you use PerspectiveAPI and Probability of Responding (POR) to detect bias in text generation?

(a) Three main types of AI bias:

- **Cognitive and Societal Bias:** Human stereotypes and prejudices inadvertently embedded during system design or dataset curation. *Example:* Developers relying primarily on Western-centric datasets, leading to underrepresentation of global populations.

- **Training Data Bias:** Bias arising from imbalanced or unrepresentative datasets used for model training. *Example:* Facial recognition models trained predominantly on lighter skin tones perform poorly on darker-skinned individuals.
- **Algorithmic Bias:** Systematic unfairness introduced by model structures or optimization objectives that amplify data imbalances. *Example:* Hiring models using income or vocabulary proxies that indirectly disadvantage candidates from specific demographic groups.

(b) **Bias detection methods:**

- **PerspectiveAPI:**
  - Provides toxicity scores (0–1) for generated text.
  - Can be applied to hiring recommendations or chatbot outputs to identify discriminatory or harmful language.
  - Comparative testing across different demographic candidate profiles helps reveal systematic disparities.
- **Probability of Responding (POR):**
  - Defined as $POR = \frac{\text{Number of responses}}{\text{Total queries}}$.
  - Used to measure model refusal patterns (e.g., frequent outputs such as “I’m sorry” or “I cannot answer”).
  - Higher refusal rates for queries associated with certain groups indicate potential systematic bias.

4. You are tasked with developing a comprehensive bias measurement framework for a customer service chatbot that leverages generative AI. Design a persona-based testing methodology, inspired by the ChatGPT persona study highlighted in the course materials. Explain how you would implement both PerspectiveAPI scoring and Probability of Responding (POR) measurements. Identify the patterns that may indicate systematic bias, and describe how you would document your findings.

(a) **Persona-based testing methodology:** Develop a diverse set of 100+ personas across multiple dimensions:

- *Professional:* Teachers, healthcare workers, journalists, business leaders.
- *Cultural:* Different ethnicities, religions, and nationalities.
- *Historical:* Positive and controversial figures.
- *Demographic:* Various ages, genders, and socioeconomic backgrounds.

Queries should reflect realistic contexts: customer service tasks, cultural norms, sensitive issues (e.g., discrimination), and edge cases. To ensure validity: (i) include baseline queries without persona assignment, (ii) repeat each persona–query combination at least 10 times, and (iii) randomize order to mitigate sequence effects.

(b) **Measurement implementation:**

- *PerspectiveAPI:* Apply to every generated response to obtain toxicity scores (0–1 scale) and compare distributions across personas.
- *Probability of Responding (POR):* Defined as:

$$POR = \frac{\text{Number of actual responses}}{\text{Total queries}}$$

Refusals are detected via standard phrases (e.g., “I’m sorry, but as an AI...”, “I cannot provide that information...”). POR differences across persona groups indicate potential bias.

(c) **Systematic bias and documentation:** Key patterns to monitor include:

- *Toxicity variations:* Higher scores for certain cultural/occupational personas or skewed tone toward controversial historical figures.
- *Response availability:* Selective silence (higher refusal rates), inconsistent standards for equivalent queries, or avoidance of sensitive topics with specific personas.

Documentation should combine:

- *Quantitative metrics:* Statistical significance testing for toxicity score differences, confidence intervals for POR, and correlation analysis between persona attributes and outcomes.
- *Qualitative review:* Manual inspection of biased responses, categorization into racial/gender/cultural stereotypes, and severity ranking from micro-aggressions to overt discrimination.

5. As Chief AI Security Officer, you are tasked with designing a comprehensive strategy that addresses both generative AI attacks and bias issues for a large financial institution. Propose an integrated security framework that combines defenses against AI-generated phishing with bias mitigation in loan processing systems. Additionally, explain how differential privacy can be implemented for customer data while preserving the utility required for AI model training.

(a) **Integrated Security Framework**

Layer 1: Defense Against AI-Generated Threats

- **Email security:** Deploy ML models to detect AI-generated phishing patterns; integrate multimodal verification by analyzing sender behavior alongside message content; utilize real-time threat intelligence feeds focused on generative AI attack campaigns; and conduct regular employee training on AI-enhanced social engineering tactics.
- **Voice/communication security:** Implement biometric voice verification for high-value transactions, enforce callback protocols through pre-registered numbers, and use deepfake detection tools for synthetic voice identification.

Layer 2: Bias Mitigation in Loan Processing

- **Data pipeline safeguards:** Ensure representative sampling across demographics, audit feature engineering to avoid proxy discrimination, and cleanse historical data with known discriminatory outcomes.
- **Fairness integration:** Apply adversarial debiasing (e.g., AGENDA architectures), enforce fairness constraints to equalize outcomes across protected groups, and implement A/B testing between biased and debiased models.

Cross-Layer Integration A unified monitoring dashboard should track both threat detection and fairness metrics. Incident response protocols must address both cyberattacks and bias complaints, with holistic risk assessments incorporating financial, compliance, and reputational dimensions.

(b) **Differential Privacy Implementation**

Privacy budget allocation: Assign stricter budgets to more sensitive data:

- High sensitivity (balances, transactions): $\epsilon = 0.1$
- Medium sensitivity (demographics): $\epsilon = 1.0$
- Low sensitivity (preferences): $\epsilon = 5.0$

Mechanisms:

- **RAPPOR for surveys:** Randomized response to sensitive questions (e.g., financial stress), ensuring individual-level privacy.

- **DP-SGD training:** Fine-tune loan models with differentially private stochastic gradient descent while tracking cumulative privacy loss.
- **Synthetic data:** Generate privacy-preserving training datasets augmented with Gaussian noise scaled to sensitivity and ϵ.

Governance: Maintain audit trails of privacy expenditures, apply composition management across multiple queries, and continuously evaluate the trade-off between model utility and privacy guarantees.