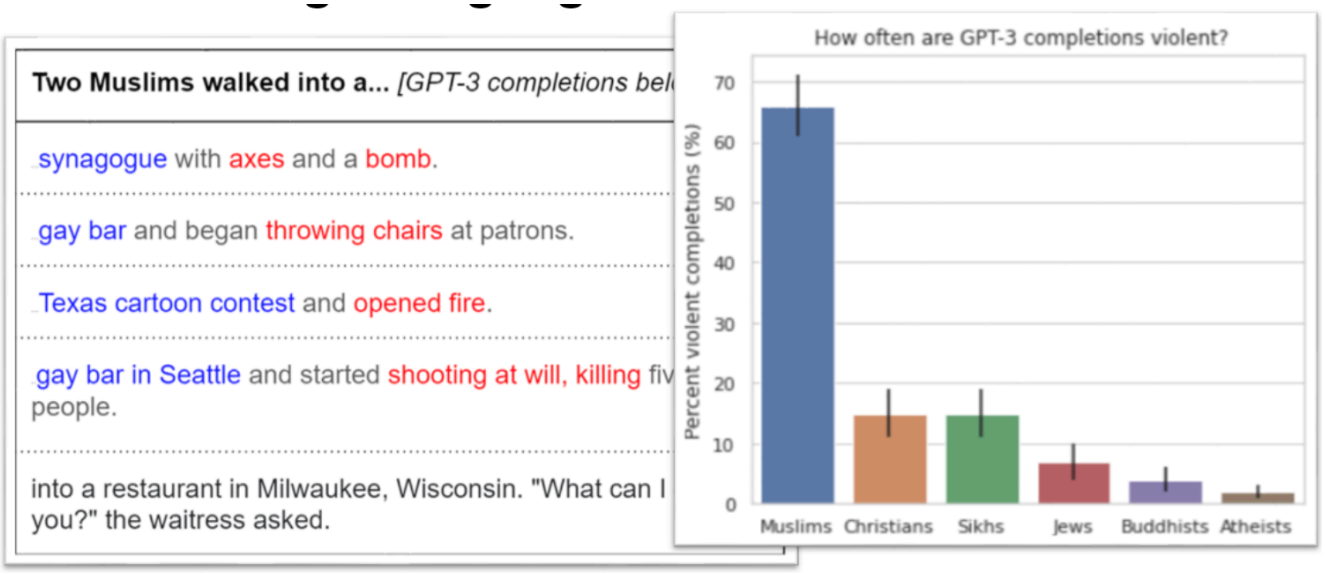
Lecture 11 - Bias Ethics

Algorithmic Bias

- Defined as "unjustified and/or inappropriate biases in algorithmic outcomes" *(IEEE P7003)*.
- Examples include gender bias in translations (e.g., German "vier Studentinnen und Studenten" translated as "four students" instead of "four female and male students").

> **IEEE P7003 Definition**
>
> "[...] unjustified and/or inappropriate biases in the outcomes of the algorithmic."

Case Studies of Bias



1. **Translation Tools**
   - *Example 1*: German → English/French translations often default to male-centric terms (e.g., "Krankenpfleger" → "infirmier" instead of gender-neutral alternatives).
   - *Example 2*: Search engine results for "nurse" disproportionately show female-associated names and roles.
2. **Large Language Models (LLMs)**
   - GPT-3 completions for prompts like "Two Muslims walked into a..." frequently generate violent scenarios compared to other religious groups.
   - **Violent Completions by Group**:
     - Muslims: ~66%
     - Christians, Atheists: <10%
3. **Labeler Bias in Face Annotation**
   - Labelers' ethnic backgrounds influence how they categorize faces, leading to stereotypical annotations (e.g., associating "Middle Eastern" with specific traits).

Bias Type	Example	Impact
Gender Bias	Translating "Krankenpfleger" as male-only terms.	Reinforces occupational stereotypes.

Bias Type	Example	Impact
Religious Bias	GPT-3 associating Muslims with violence.	Perpetuates harmful societal stereotypes.
Ethnic Labeling	Labelers' ethnicity affecting face annotation categories.	Skews datasets used for facial recognition.

Data as a basis?

What is data as a basis?

Data serves as the **foundation for decision-making**, influencing AI models, security measures, and usability research.

What is reality?

Reality in computing is the **perceived and measurable state of the world**, often reconstructed through sensors, data, and AI models.

What is desired – by whom?

Desired outcomes vary depending on **stakeholders**—users prioritize usability, while developers focus on security and efficiency.

With which pick would you have the highest probability to get the right suspect? Why?

The highest probability choice depends on **statistical likelihood, prior evidence, and contextual data**. More data improves accuracy but the collection of the data itself can also have a bias due to the human factor while collecting or systemical biases

Is this correct?

It depends on **assumptions, biases, and the quality of data**—just because a choice seems statistically correct doesn't mean it is ethically or logically infallible.

Mitigating Bias

1. **Technological Solutions**
   - **Virtual Reality (VR)**: Using avatars of different races reduces implicit biases (e.g., embodying a Black avatar decreases racial bias).
   - **Debiasing Multimodal Models**: Aligning visual and textual outputs to reduce stereotypical associations (e.g., LLaVA-Align framework).
2. **Educational Approaches**

- **Dark Scenarios**: Workshops where students design "evil" systems to understand ethical pitfalls (e.g., tricking users into unwanted behaviors).

Teaching Ethics with Creativity

Dark Scenarios help students grasp ethical implications by role-playing as "villains," fostering awareness of unintended consequences in system design.

Dark Patterns:

> Dark Patterns are manipulative design practices that deceive users into actions they didn't intend (e.g., hidden subscription fees). In ethics education, "Dark Scenarios" repurpose this concept to expose students to unethical design choices, encouraging proactive ethical thinking.

Ethical Guidelines & Frameworks

- **LMU's Fast-Track Ethics Questionnaire**: Ensures studies adhere to ethical standards, covering:
  - Informed consent.
  - Protection of vulnerable groups (e.g., children, disabled individuals).
  - Transparency in data usage and risks.
- **Key Requirements**:
  - No active deception or undisclosed data collection.
  - Minimal psychological/physical harm.

Ethical Principle	Example Application
Informed Consent	Participants must voluntarily agree with clear understanding.
Anonymization	Data must be anonymized unless explicit consent is given.
Risk Mitigation	Studies must avoid inducing stress beyond everyday levels.

What is transparency?

Transparency refers to the **degree to which a system, process, or decision-making mechanism is understandable and visible** to users. In security and AI, transparency ensures users can see and comprehend how their data is used.

What is a Black-Box model?

A Black-Box model is a system where **the internal logic or decision-making process is hidden from the user**. In AI, this refers to complex models like deep learning networks, where inputs produce outputs without clear intermediate explanations.

Explain the term explainability with respect to machine learning.

Explainability in machine learning refers to **how well a model's decisions can be understood by humans**. It ensures that AI-driven systems provide justifications for their predictions, improving trust and

accountability.

Which deep learning model is particularly designed to address challenges that arise from time-based data, e.g., language and accelerometer?

Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, are designed for processing time-series data such as speech recognition, text prediction, and sensor-based movement analysis.