Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 2: Overview (AI Ethics)

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

February 21, 2025

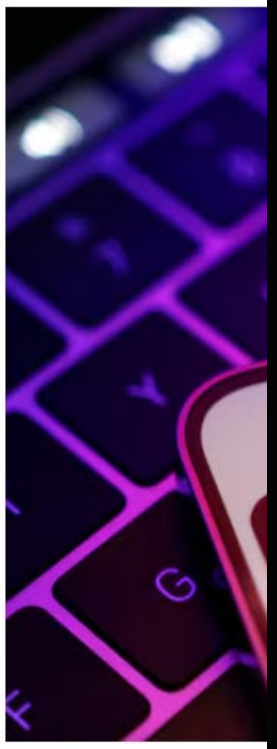
Reuters

World Business

Musk's xA
rival Chat

By Reuters

February 18, 2025 10:55 PM



xAI and Grok logos are seen in this
Rights

Feb 18 (Reuters) - Elon M
of its chatbot, as it looks t
OpenAI, and Alphabet's (

Grok-3 debut comes at a
powerful open-source mo

xAI Grok API Blog About Careers

February 19, 2025

Grok 3 Beta — The Age of Reason

We are thrilled to unveil ar
advanced model yet, blen
extensive pretraining kno

Next-Generation Intelligence from xAI

We are pleased to introduce Grok 3, our most advanced model yet: blending strong reasonin
knowledge. Trained on our Colossus supercluster with 10x the compute of previous state-of
significant improvements in reasoning, mathematics, coding, world knowledge, and instructi
reasoning capabilities, refined through large scale reinforcement learning, allow it to think fo
errors, exploring alternatives, and delivering accurate answers. Grok 3 has leading performar
benchmarks and real-world user preferences, achieving an Elo score of 1402 in the Chatbot
Grok 3 mini, which represents a new frontier in cost-efficient reasoning. Both models are still in training and will evolve rapidly
with your feedback. We are rolling out Grok 3 to users in the coming days, along with an early preview of its reasoning

NEWS

XAI's Grok-3 highlights openness and transparency concerns

THE ECONOMIC TIMES | News

English Edition | Today's ePaper

My Watchlist Sign In

Special Offer on ETPrime

Home BUDGET'25 ETPrime Markets Market Data News Industry Rise Politics Wealth MF Tech Careers Opinion NRI Panache Videos

India Web Stories Economy Politics Newsblogs Elections Defence International More

Business News › News › International › Global Trends › Is Grok-3 putting software engineer jobs at risk? Musk's xAI team says AI chatbot saved hundreds of hours on coding

Bears Growl Sensex slips 500 pts, Nifty below 22,800; auto, pharma top sectoral losers

Is Grok-3 putting software engineer jobs at risk? Musk's xAI team says AI chatbot saved hundreds of hours on coding

ET Online • Last Updated: 19 February, 2

SEARCH FORTUNE SIGN IN Subscr

NEWSLETTERS · EYE ON AI

AI security risks are in the spotlight—but hackers s models are still alarmingly easy to attack

BY SHARON GOLDMAN

February 19, 2025 at 4:31 AM GMT+8

Home / Innovation / Artificial Intelligence

Yikes: Jailbroken Grok 3 can be made to say and reveal just about anything

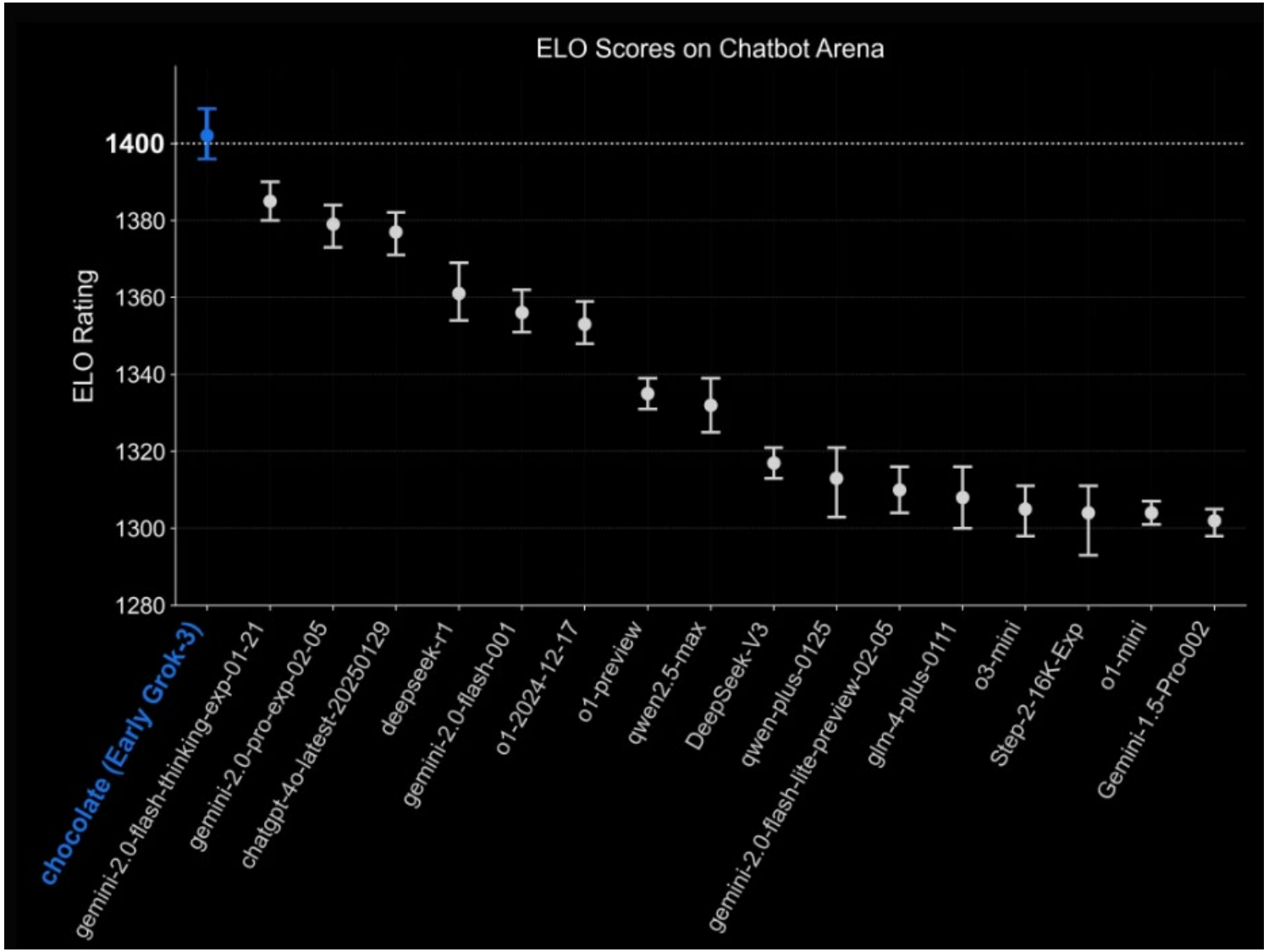
A red team got xAI's latest model to reveal its system prompt, provide instructions for making a bomb, and worse. Much worse.

Written by **Radhika Rajkumar**, Editor

Feb. 19, 2025 at 9:32 a.m. PT

Benchmark	Grok 3 Beta	Grok 3 mini Beta	GPT-4o	Gemini 2.0 Pro	DeepSeek-V3	Claude 3.5 Sonnet
AIME'24	52.2%	39.7%	9.3%	—	39.2%	16.0%
GPQA	75.4%	66.2%	53.6%	64.7%	59.1%	65.0%
LCB	57.0%	41.5%	32.3%	36.0%	33.1%	40.2%
MMLU-pro	79.9%	78.9%	72.6%	79.1%	75.9%	78.0%
LOFT (128k)	83.3%	83.1%	78.0%	75.6%	—	69.9%
SimpleQA	43.6%	21.7%	38.2%	44.3%	24.9%	28.4%
MMMU	73.2%	69.4%	69.1%	72.7%	—	70.4%
EgoSchema	74.5%	74.3%	72.2%	71.9%	—	—

https://x.ai/blog/grok-3



https://x.ai/blog/grok-3

Goals of this lecture

- After this lecture, students will learn:
  - Principles of Ethical AI
  - Ethical issues in biomedical research

XPRIZE Video: AI for Good - AI and Medicine

- Artificial intelligence will change lives in many ways. Already, AI solutions are being deployed and having significant impact in healthcare. Daniel Kraft, MD, Chair for Medicine, Singularity University, shares his expert views on how significant this technology will be in finding the right diagnosis and therapies and shifting the ‘practice’ of medicine to the real ‘science’ of medicine. (May 3, 2017)

XPRIZE Video: AI for Good - Ethics in AI

- Address AI from an ethics, safety, moral and privacy rights perspective and the need for a guiding ethical framework and code of conduct to create a foundation for the design, production and use of AI. (May 27, 2017)

Definitions

- **Ethics**: The systematic set of principles or guidelines that govern conduct within a particular context, often derived from philosophical theories and professional standards. It deals with questions of what is right and wrong, and what individuals ought to do in various situations.
- **Values**: Individual or collective beliefs about what is important, desirable, and worthwhile. They reflect personal or cultural priorities and inform decisions and behaviors.
- **Morality**: The principles or rules of behavior that individuals or societies believe are right or wrong. It is more focused on practical, everyday conduct and often carries a strong emotional and social component.

Etymology of Ethics

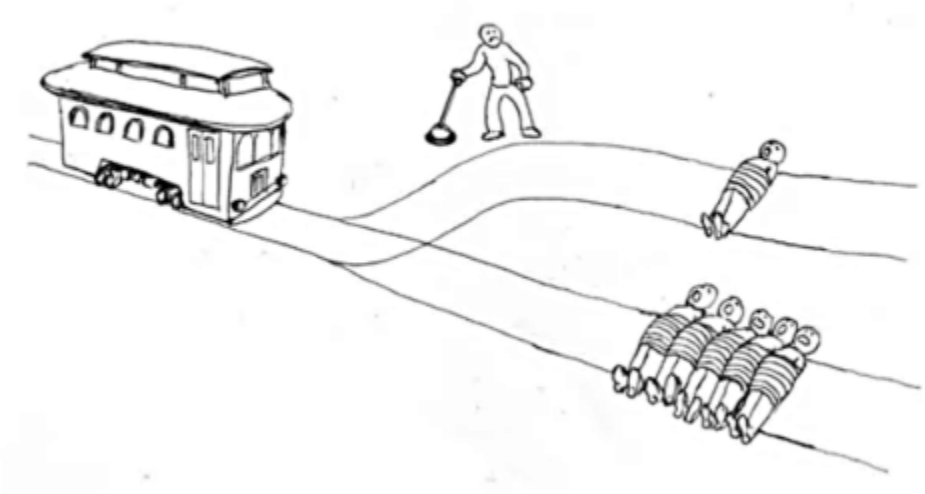
- Etymologically, the English word “ethics” (ethica in Latin) can be traced back to the ancient Greek noun, (ethos), which denotes a “habit” or “custom”.
- Ethics is a practice discipline that refers to human action with the purpose of being morally good.
- In Chinese, “Lun” means “Order” which represents relationships; “Li” means “Rule” which represents principles.
- “Morality” is more subjective. “Ethics” is more objective.

Introduction to Ethics

- **Definition:** Ethics is the branch of philosophy that deals with questions about what is morally right and wrong, fair and unfair, good and bad.
- **Purpose**: To guide human behavior, ensuring individuals and organizations act in a morally responsible way.
- **Healthcare Ethics**: Ensures that professionals make decisions in the best interests of patients (e.g., informed consent, confidentiality).
- **Technology Ethics**: Ethical concerns regarding data privacy, AI development, and the impact of technology on society.

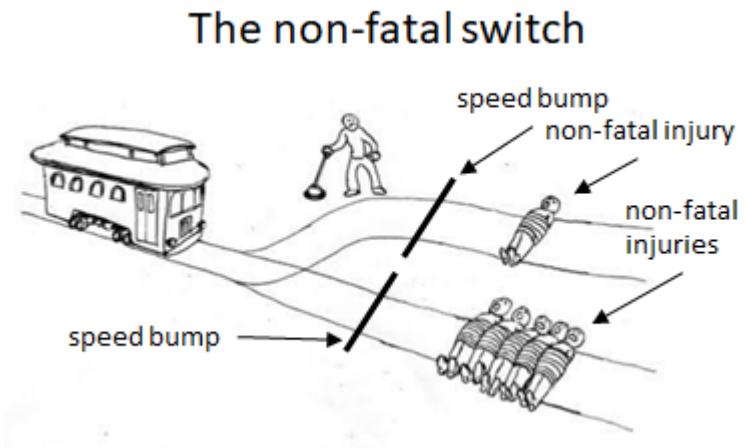
What would you do?

- Trolly Problem



What would you do?

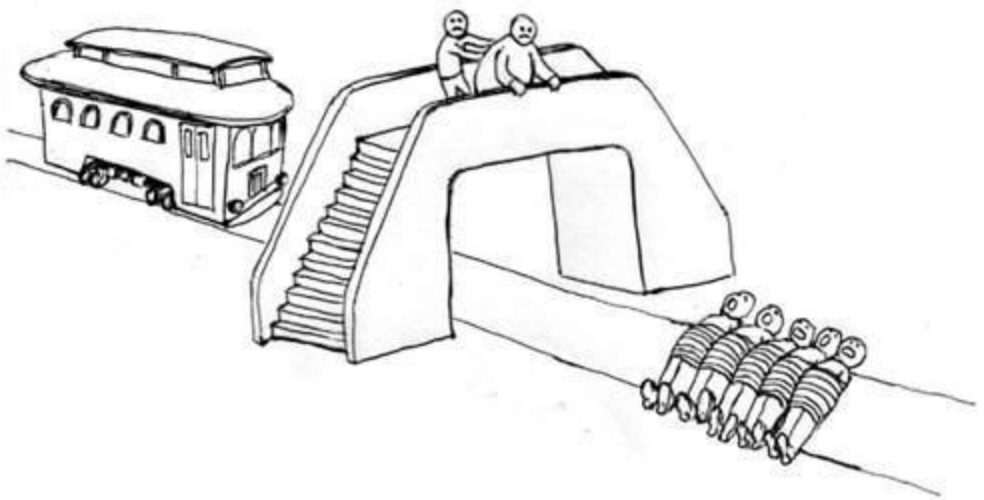
- V1



https://themindcollection.com/trolley-problem-meme-variations/

What would you do?

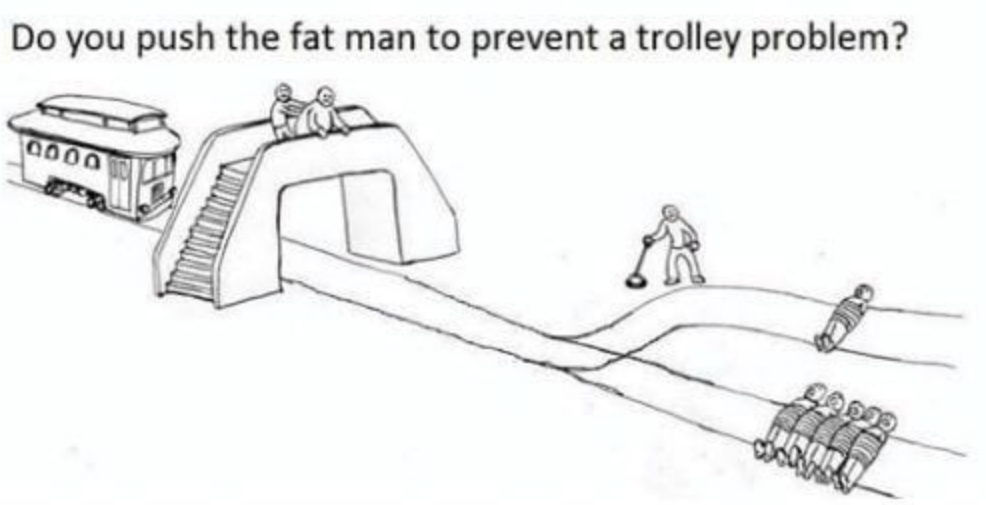
- V2



https://themindcollection.com/trolley-problem-meme-variations/

What would you do?

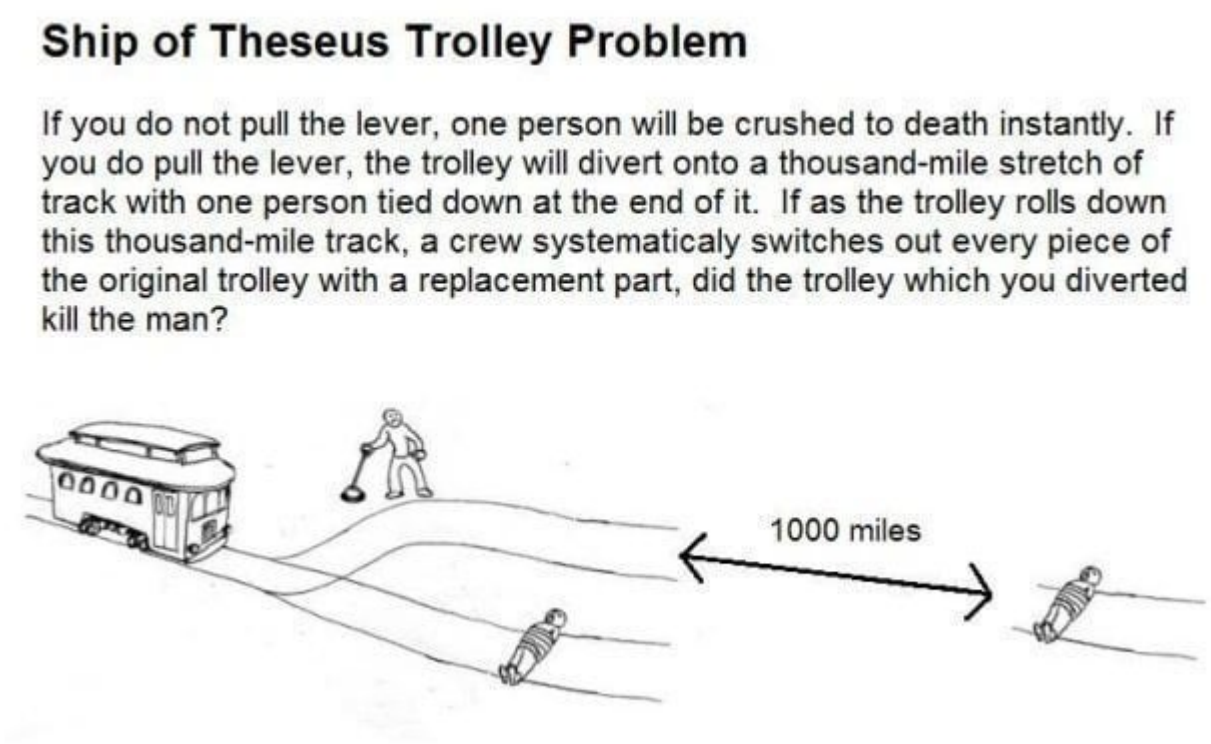
- V3



https://themindcollection.com/trolley-problem-meme-variations/

What would you do?

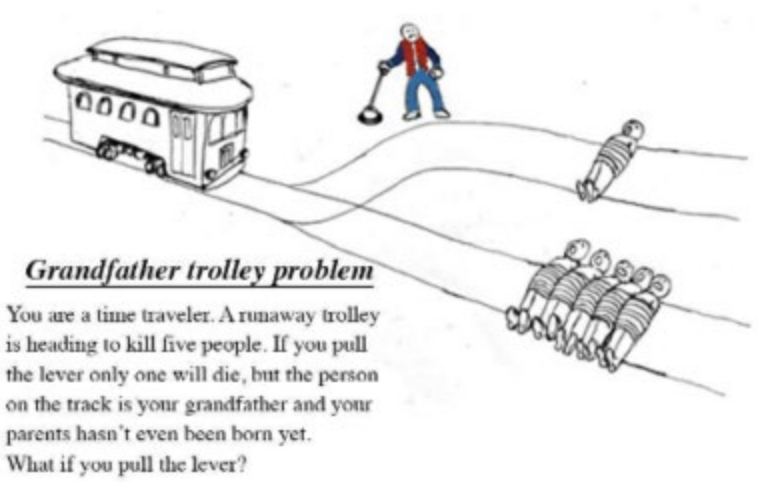
- V4



https://themindcollection.com/trolley-problem-meme-variations/

What would you do?

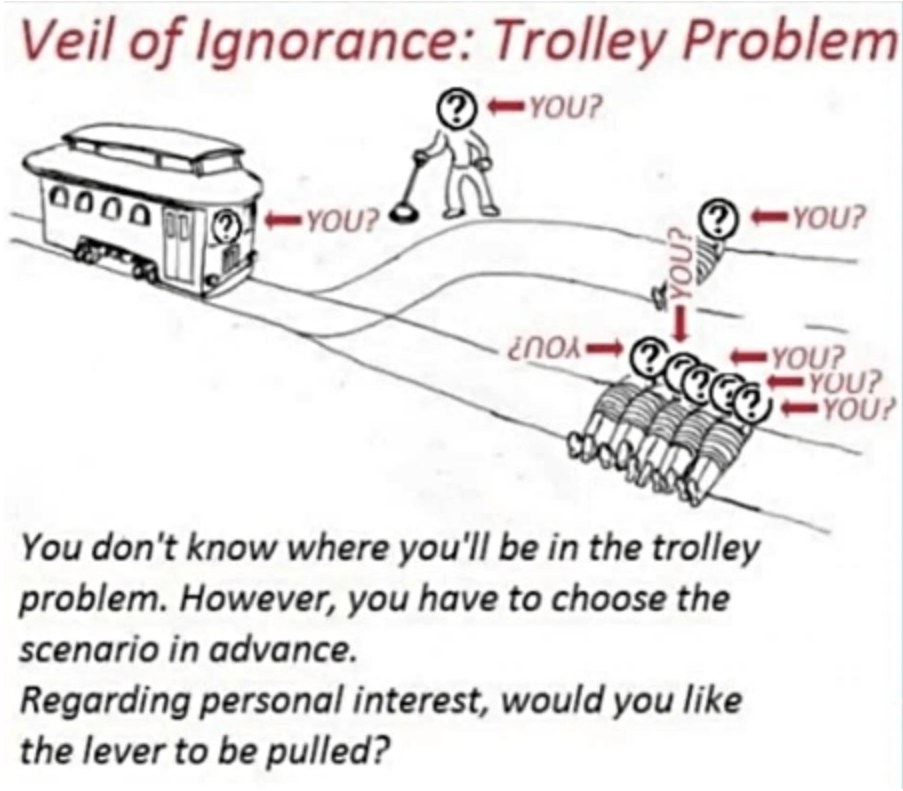
- V5



https://themindcollection.com/trolley-problem-meme-variations/

What would you do?

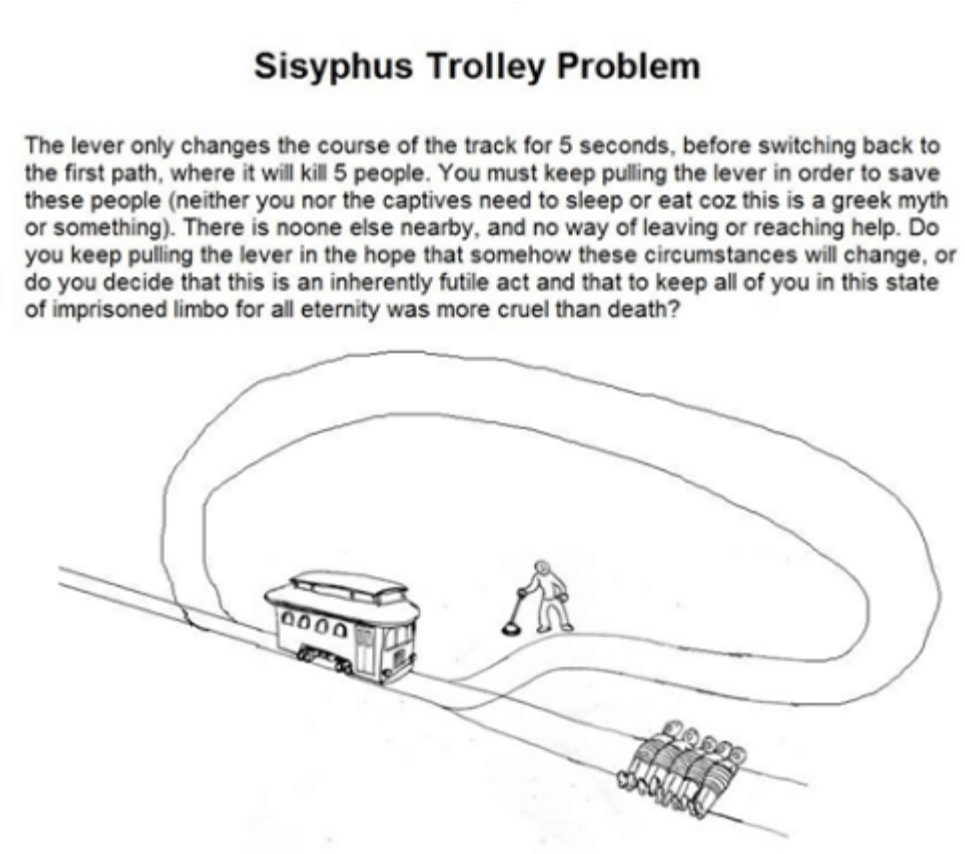
- V6



https://themindcollection.com/trolley-problem-meme-variations/

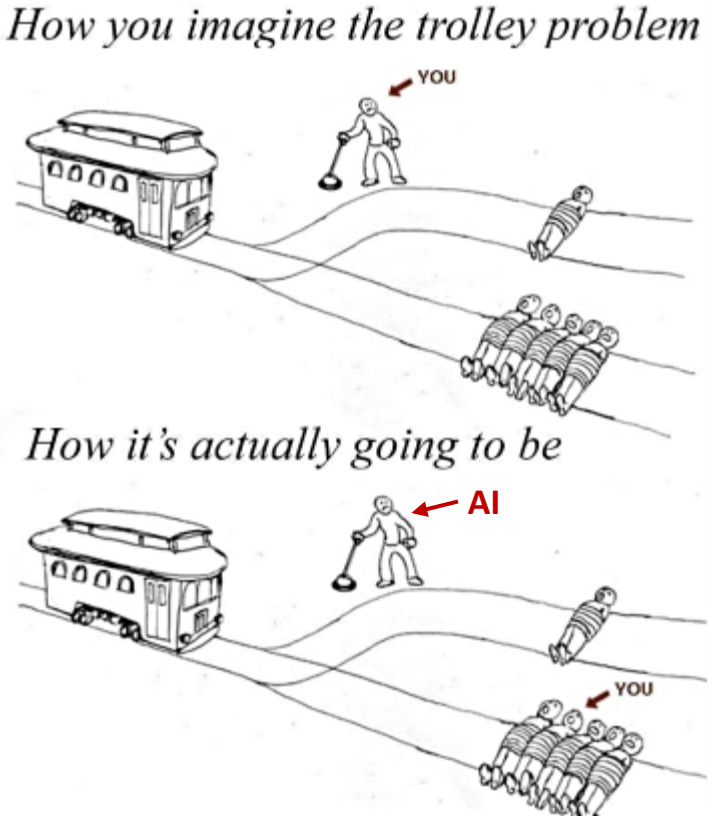
What would you do?

- V7



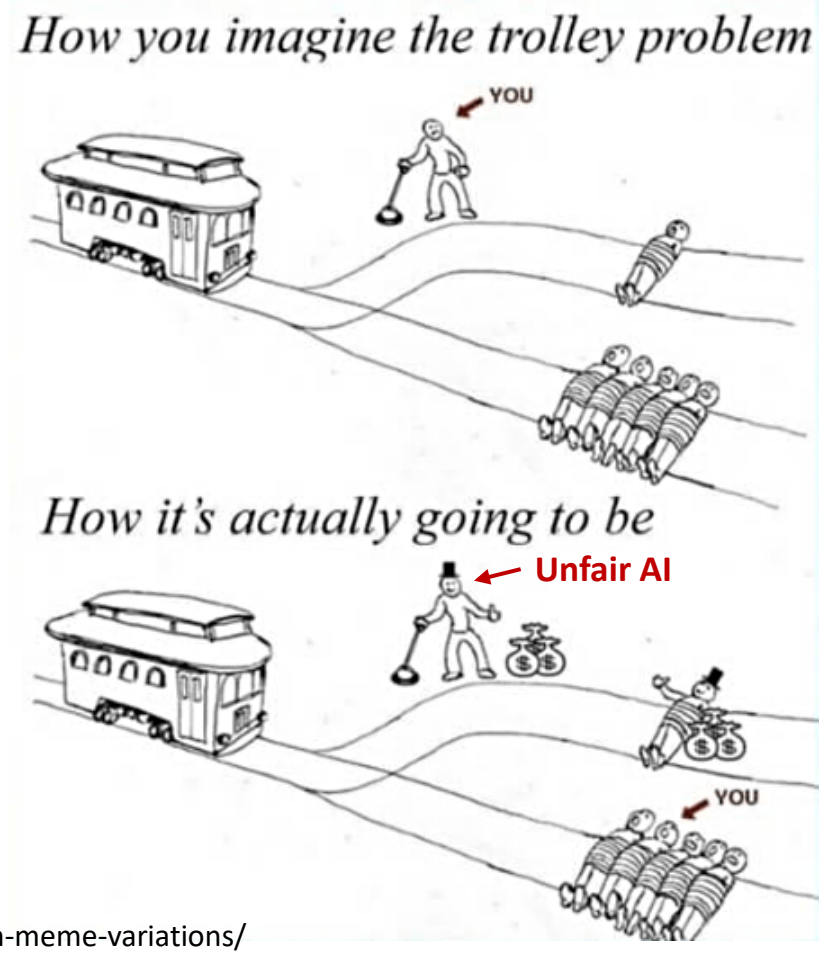
https://themindcollection.com/trolley-problem-meme-variations/

Actual Trolly Problem in the Age of AI



https://themindcollection.com/trolley-problem-meme-variations/

Trolly Problem Shows Challenges of AI Ethics



https://themindcollection.com/trolley-problem-meme-variations/

What should the AI (self-driving car) do?

- **Scenario:** Cargo fell off the truck ahead, and the brakes are insufficient to avoid a collision. The **self-driving car** faces three options: first, turn left and collide with an SUV; second, turn right and collide with a cyclist; third, go straight and crash into an obstacle ahead.
- Strategies:
  - 1. Prioritize self-safety.
  - 2. Minimize overall damage.
  - 3. Additional information.
  - 4. Random strategy.

Recap of AI Models

Markov Models	• Markov models are probabilistic models that capture dependencies between events.
Neural Network Models	• Neural network models are computational models inspired by the human brain.
Generative AI	• Generative AI combines the concepts of Markov and neural network models. • It leverages the probabilistic nature of Markov models and the learning ability of neural network models to generate new data or content.

Brief History of AI Development

17th Century	Pascal and Leibniz	Ideas of Intelligent Machines
1920s	Charles Babage	1st “Computing Machine”
1950	Alan Turing	Proposed “Turing Test”
1955-1956	John McCarthy	Organized the Dartmouth Conference, proposed the concept of AI
1997	IBM	Deep Blue (Won Chess World Champion)
1990s	Vapnik and Chervonenkis	SVM
1998	Yann LeCun	LeNet-5 -> CNN
2006	Hinton	Neural networks
2012	Hinton et al.	AlexNet
2014	Ian Goodfellow	Generative adversarial network (GAN)
2015	Kaiming He et al.	ResNet
2017	Google DeepMind	AlphaGo (Won Go Chess World Champion)
2018	Google AI	Bidirectional Encoder Representations from Transformers (BERT)
2021	Google DeepMind	AlphaFold
2022	OpenAI	ChatGPT-3.5 (ChatGPT based on GPT-3.5)
2023	OpenAI	ChatGPT-4 (ChatGPT based on GPT-4)

Group Discussion

- In a group of 2-3, discuss the following questions:
  - What should an ethical AI system look like?
  - What properties and principles should it have?
  - Can you give several examples of ethical AI systems?
  - Can you give several examples of unethical AI systems?

Responsible AI ≈ Ethical AI

- **Transparency**: AI systems should be clear and understandable
- **Fairness**: AI systems should be unbiased and not discriminate against any group of people
- **Privacy**: AI systems should protect people's personal information
- **Accountability**: Organizations should be held accountable for how they use AI
- **Non-maleficence**: AI systems should not harm individuals, society, or the environment
- **Inclusiveness**: AI systems should engage with diverse perspectives

UNESCO	IEEE	IBM
UNESCO ethical recommendations are based on specific core values such as human dignity and rights, promoting peace, and care for the environment. Based on these values, UNESCO specifies ten principles: 1. Proportionality and Do No Harm 2. Safety and Security 3. Right to Privacy and Data Protection 4. Multistakeholder and Adaptive Governance & Collaboration 5. Responsibility and Accountability 6. Transparency and Explainability 7. Human Oversight and Determination 8. Sustainability 9. Awareness and Literacy 10. Fairness and Non-discrimination [38]	The IEEE Standards Association (SA) has established a Global Initiative on the Ethics of Autonomous and Intelligent Systems. The IEEE approach is established on eight fundamental principles: 1. Human Rights, 2. Well-being, 3. Data Agency, 4. Effectiveness, 5. Transparency, 6. Accountability, 7. Awareness of Misuse, and 8. Competence [39]	IBM proposes three guiding values for AI: 1. The purpose of AI is to augment human intelligence, 2. Data and insights belong to their creator, and 3. Technology must be transparent and explainable. Leveraging insights from the 1979 Belmont Report, IBM defines three overarching principles for AI: 1. Respect for persons, 2. Beneficence, and 3. Justice, i.e., burdens and benefits may be distributed either by: a. Equal share, a. Individual need, a. Individual effort, a. Societal contribution, or a. Merit [40]

Table 1. Ethical Principles Statements from selected organizations

Kirova VD, Ku CS, Laracy JR, Marlowe TJ. The ethics of artificial intelligence in the era of generative AI. Journal of Systemics, Cybernetics and Informatics. 2023 Dec;21(4):42-50.

Transparent vs Explainable

- **Scenario**: A financial institution uses a decision tree model to approve loans and provides explanations like “Loan denied because the applicant’s credit score is below 700 and their debt-to-income ratio is above 40%.”
- **Explanation**: The model gives clear reasons for each decision based on the decision tree’s structure.
- **Lack of Transparency**: The proprietary data and criteria used to build the decision tree are not shared, preventing a full understanding of the data sources, potential biases, and overall decision-making process.

Transparent vs Explainable

- **Scenario**: An image recognition system using a deep neural network is fully open-source, with all model weights, architectures, and training processes documented and available.
- **Transparency**: Every aspect of the model's development, including data, training code, and hyperparameters, is openly accessible.
- **Lack of Explanation**: Despite the transparency, the decisions made by the neural network (e.g., why it classified an image as a cat) are not easily interpretable or understandable to most users due to the complexity of the model.

Transparent vs Explainable

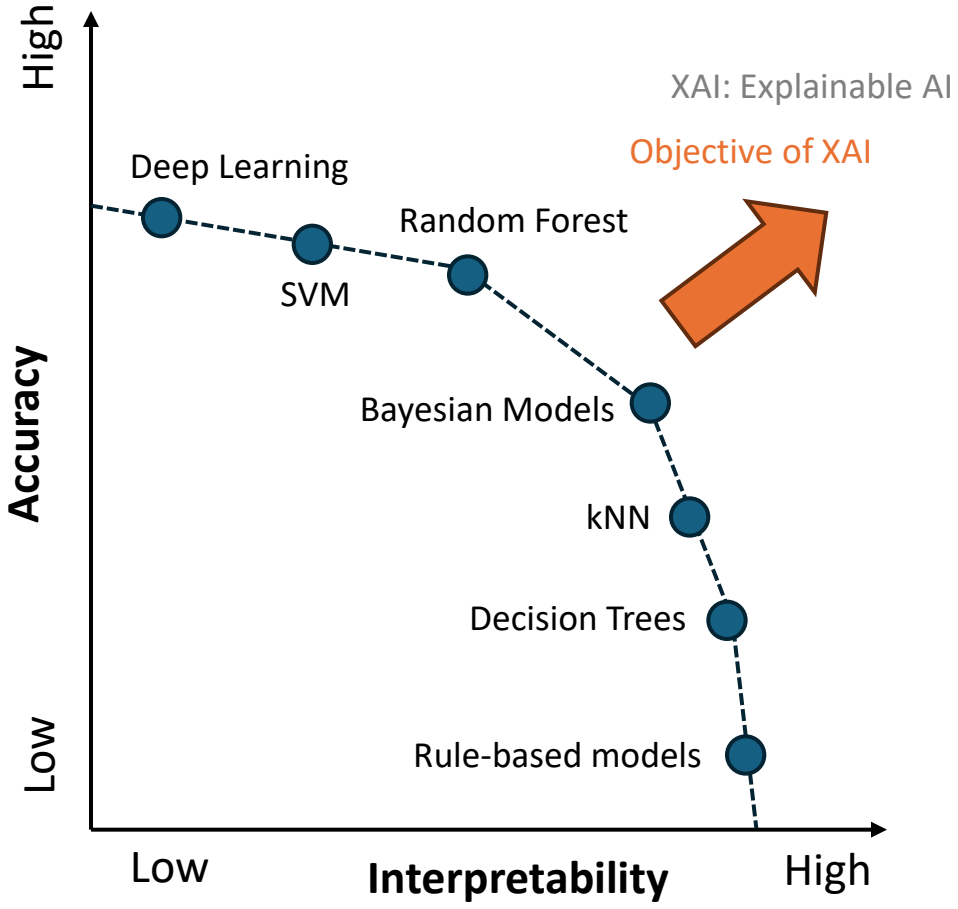
- **Scenario**: A machine learning pipeline used for predicting stock prices is entirely transparent, with all preprocessing steps, feature engineering, and model selection processes documented and shared publicly.
- **Transparency**: The pipeline's design, implementation details, and data sources are fully disclosed.
- **Lack of Explanation**: The final model's predictions (e.g., why a specific stock is predicted to rise) are not easily explainable to users because the model uses complex algorithms like ensemble methods or deep learning, making it hard to attribute specific factors to individual predictions.

Transparent vs Explainable

- **Scenario**: A rule-based expert system used in healthcare provides specific reasons for its recommendation (e.g., “Patient should take medication X because their blood pressure is above 140/90 and they have a history of heat disease”).
- **Explanation**: They system provides clear, understandable explanations for its decisions based on predefined rules.
- **Lack of Transparency**: The underlying logic and how these rules were derived are not disclosed to users. The development process and criteria for rule creation are hidden, making it difficult to understand the overall system’s logic and potential biases.

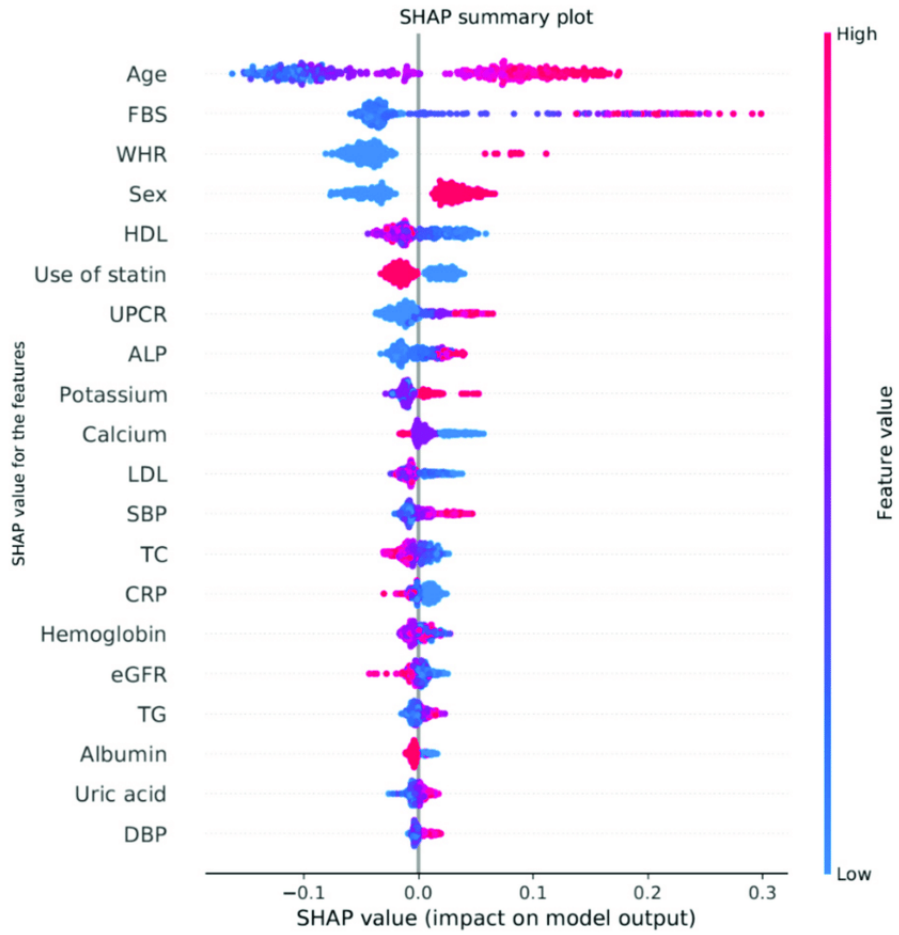
Tradeoffs

- There is often a tradeoff between the performance and explainability of AI
- There is often a tradeoff between the transparency and privacy of AI
- Manual human processes are rarely transparent, unbiased, or explainable



Transparency and Explainability Solutions

- SHAP (Shapley Additive exPlanations)
  - ➢Pros
    - Model agnostic
    - Local explanation
    - Unified framework
    - Feature importance
  - ➢Cons
    - Missing interaction effects



Group Discussion

- In a group of 2-3, discuss the following questions:
  - How to handle the deep integration of human and AI, or in other words, who should be responsible for making decisions?
  - How can the issues caused by the opacity and lack of interpretability in AI systems be addressed?
  - If an accident occurs due to a problem with an AI system, who should be responsible for the accident?

Simpson's Paradox

- Berkley gender bias in the 1970s.
- All departments admitted men at higher rates.
- The university admitted women at higher rates.
- Who is correct? Does gender bias exist?

	Women		Men
Dept. A	0/1	<	50/100
Dept. B	70/100	<	1/1
Total	70/101	>	51/101

Medical Outcomes

- Are white people healthier than black people?
- “Blacks have 26.3% more chronic illnesses than Whites” ($P < 0.001$)
- Removing algorithm bias by substituting “healthier whites” with “less healthy blacks” until the “marginal patient is equally healthy”.
- This creates substantial disparities in health screening.

Policing

- Predictive policing uses AI to forecast crime likelihood and proactively police areas.
- Data is typically drawn from prior-arrest databases.
- This creates a feedback loop.
- Potential bias in arrests.

Definition of Bias

- 1. prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.
- 2. a systematic distortion of a statistical result due to a factor not allowed for in its derivation.

Bias from a Legal Perspective

- Protected class
- Sensitive characteristics
- Disparate outcome
- Predictive parity

Bias from a Data Perspective

- **Sampling Bias**: Occurs when the sample data is not representative of the population intended to be analyzed
- **Survivorship Bias**: Focusing on data from “survivors” of a process while ignoring those that did not make it through
- **Data Collection Bias**: Bias introduced during the data collection process due to inconsistent or flawed methodologies
- **Reporting Bias**: Arises when only certain outcomes or data points are reported, often those that support a particular hypothesis

Bias from a Data Perspective

- **Social Desirability Bias**: Respondents provide answers that are more socially acceptable than their true thoughts or behaviors
- **Publication Bias**: Studies with significant/positive results are more likely published, skewing perception of research outcome
- **Historical Bias**: Results from biases present in historical data that are perpetuated in current models
- **Algorithmic Bias**: Bias introduced by the design and functioning of algorithm itself

Sampling Bias

- **Scenario**: A tech company is developing an AI-based facial recognition system for gender and uses a dataset predominantly composed of images from public figures and celebrities.
- **Bias**: This dataset is likely to underrepresent older individuals, people of varying attractiveness, and ethnic minorities. As a result, the AI model trained on this dataset may perform poorly when recognizing faces outside these demographic groups.
- **Implication**: The facial recognition system may exhibit significant inaccuracies and higher error rates for underrepresented groups, leading to biased and unreliable results in practical applications.

Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency 2018 Jan 21 (pp. 77-91). PMLR.

Survivorship Bias

- **Scenario**: During WWII, returning aircraft were analyzed for where to add amor. They observed damage on wings and fuselage, and thus suggested reinforcing these areas.
- **Bias**: This analysis only included planes that survived and returned from missions. The missing data were from planes that were shot down and did not return, which might have been hit critical areas like the engines.
- **Implication**: Focusing on the surviving aircraft led to incorrect conclusions. The real vulnerabilities were in the parts that, when hit, caused planes to be lost.

Mangel M, Samaniego FJ. Abraham Wald's work on aircraft survivability. Journal of the American Statistical Association. 1984 Jun 1;79(386):259-67.

Social Desirability Bias

- **Scenario**: A tech company is developing a sentiment analysis AI to gauge public opinion on sensitive topics by collecting survey data on controversial issues like racial discrimination or political views.
- **Bias**: Respondents may provide socially acceptable answers rather than their true opinions to avoid judgement or backlash, leading to social desirability bias.
- **Implication**: people are going to report what they think is the right answer as opposed to what they truly believe, especially in something like customer survey or sentiment analysis.

Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Quality & quantity. 2013 Jun;47(4):2025-47.

Historical Bias

- **Scenario**: A company develops a hiring algorithm designed to screen resumes and predict job performance based on historical hiring data.
- **Bias**: Training data predominantly includes resumes of employees who were hired and performed well in the past, which may reflect historical biases favoring certain demographics. The algorithm may favor resumes that resemble those of historically preferred candidates, while disadvantaging equally qualified candidates from underrepresented groups.
- **Implication**: It's very difficult when you're developing a selection tool to use your existing population.

Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 conference on fairness, accountability, and transparency 2020 Jan 27 (pp. 469-481).

Algorithmic Bias

- **Scenario**: An AI system predicts the likelihood of patients developing complications after surgery, using preoperative health data
- **Bias**: The algorithm is trained on data where certain demographic groups (e.g., younger patients or those with fewer comorbidities) are overrepresented. If the model relies heavily on these characteristics, it may inaccurately predict lower risk for older patients or those with more complex medical histories, leading to under-preparation and potentially poorer outcomes.
- **Implication**: The AI system may fail to predict complications for diverse patient groups that are not like it was trained on.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018 Dec 18;169(12):866-72.

Mitigating Bias

- Ensure the sample is representative and randomly selected.
- Use validated and reliable measurement instruments.
- Train data collectors thoroughly to minimize observer bias.
- Collect data from multiple sources and contexts.
- Transparently report all data, including null and negative results.
- Regularly audit and evaluate data and algorithms for bias.
- Include diverse perspectives in the data collection and analysis process.

Case studies

- What types of data were misused? What principles of ethical AI were violated? What AI model were used? How to mitigate the risks?

Case 1	In 2018, Amazon's facial recognition system misidentified 28 lawmakers as criminals.
Case 2	In 2017, Vietnamese security company, Bkav, used 3D-printed mask to bypass iPhone's Face ID.
Case 3	In 2019, Criminals used AI technology to mimic CEO's voice.
Case 4	UK passport photo AI verification system shows bias against Black women.
Case 5	Amazon's voice assistant Alexa recommends a 10-year-old girl to use a coin to touch a socket.
Case 6	YouTube's recommendation algorithm suggests inappropriate videos to children.
Case 7	Ride-hailing platforms recommend different car models based on the user's phone brand and price.
Case 8	In 2020, Tesla's self-driving car failed to recognize a white truck, leading to an accident.
Case 9	In 2018, Uber's self-driving vehicle stroked a pedestrian at night.

Group Discussion

- In a group of 2-3, discuss the following questions:
  - There are now various methods of identity recognition, including facial recognition, voice recognition, and fingerprint recognition, with related AI technologies becoming increasingly mature. However, data such as facial photos and voice recordings are easily accessible, posing significant security risks. Can you think of some other identity recognition methods with higher security?
  - Long-term use of recommendation systems can easily lead to the “filter bubble” effect. How can users avoid falling into this “filter bubble”? Do you prefer the system to recommend the most interesting information to you, or would you prefer it to recommend information from different fields?
  - What ethical issues exist in robot care for the elderly? How to solve?
  - What ethical issues exist in Brain Computer Interface? How to solve?

Brainstorming

- “In 15 years, AI and automation will have the technological capability to replace 40% of jobs.” – Kaifu Li

Jobs will be replaced by AI in 15 years	Jobs will NOT be replaced by AI in 15 years
	• Medical researcher, artificial intelligence scientist, screenwriter, public relations expert, entrepreneur. • CEO, negotiation expert, mergers and acquisitions expert. • Oral surgeon, aircraft mechanic, chiropractor. • Geological survey cleaner. • Social worker, special education teacher, marriage counselor.

Readings for the Next Lecture

- Price WN, Cohen IG. Privacy in the age of medical big data. Nature medicine. 2019 Jan;25(1):37-43.
  - ❑ https://www.nature.com/articles/s41591-018-0272-7
- Optional
  - ❑《工程伦理》Ch.10.
  - ❑《信息科学技术伦理与道德》Chs.5&7.
  - ❑ S. Warren and L. Brandeis. The right to privacy. Harvard Law Review. 1890; V. IV, No. 5.
    - http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm

ON A PIECE OF PAPER

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"