# AI and Ethics Challenges


**New Mexico Courts, Corrections & Justice Committee**

**August 2025**

# Questions raised by AI

- AI is pervasive, integrated in other technologies
  - Often invisible
  - Ethically – is the problem the AI or the host technology?
- AI is not only about technology – it is about what humans do with it
  - How it is used, how it is perceived, how it is embedded in daily lives
  - AI Ethics discussion should consider the social context in which AI is used

# AI Hype and Fears

# AI Better than Humans?

- 1997: DeepBlue beats Kasparov at Chess (search algorithms, domain knowledge)
- 2016: AlphaGo beats Lee Sedol at Go (machine learning)


- Admiration for beauty of moves and strategies
- Sadness – loss of dominance in games where human creativity was considered essential

**Dave: "Open the pod bay doors"**
**HAL: "I'm afraid I can't do that, Dave."**

# Superintelligence and transhumanism

- **Hype Surrounding the future of AI**

- Concept repeated in popular media and public discourse
- Influential people
  - Elon Musk (2022): artificial intelligence could one day outsmart humans and endanger us, citing AI as the biggest threat to civilization
  - Hawking (2014): The development of full artificial intelligence could spell the end of the human race
- **Superintelligence**
  - Machines will surpass human intelligence
  - Machine will Master Us

# The real dangers

- Data
  - Provenance
  - Privacy and Security
- Context
  - Unintended
  - Intended
- Responsibility (legal and non)
- Human Control
- Inclusion and Participation

# Privacy and Data Protection

- ML use big data sets often inclusive of personal information
  - E.g., images gathered from street cameras, smartphone locations, social media data
  - Big data are often aggregations of heterogenous data sets
- Ethical use would imply
  - Right of privacy of individuals
  - Right to know what data are used, how, and by whom
- Challenge with respect to traditional social science research
  - One collector – one use
  - Consider social media – consent given to allow access to the platform, lack of awareness of what data collected, data moved to other domains or sold to other companies
- Research on **Differential Privacy in Algorithms**


# Manipulation, Exploitation, Vulnerable Users

- User exploitation
  - Social media users are free labor to data companies – they analyze and sell data to companies that target consumers
    - Amazon mechanical turk


  Hidden risk to democracy – not as authoritarian politics
  - But by changing the economy – we provide free data and we are guided in our decisions
  - More direct political influence
    - Cambridge Analytica – facebook data used without consent
    - Bots building targeted political messages
  - AI taking over cognitive tasks from humans ("infantilizes users")

# Manipulation, Exploitation, Vulnerable Users

- Vulnerable users – not all AI users are young healthy adults
  - Elderly easily manipulated in believing information provided by social media bots
  - Young children interacting with internet toys – provide personal and family data, manipulated in their learning process

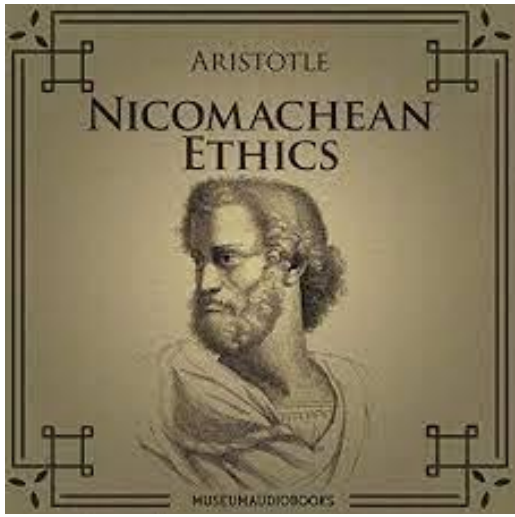
# False Information and Impact on Personal Relationships

- AI used to create hate speech and false information
  - Bots disguised as humans
    - Chatbot Tay [Microsoft]
      - Designed to engage 18-24-year-old crowd on Twitter (playful conversations)
      - Learning system – twitter users repeated offensive/racist language
      - Tay created racist speeches
    - Univ. Washington researchers created AI tool to generate fake Obama speeches
  - Unclear what is true and what is false
    - Not new (e.g., newspapers propaganda) but more diffuse, personalized, intense
  - History shows that confusion of truth used for ideological purposes
- Even in a more utopian view
  - AI creates digital companions, illusion of companionship
  - Unsettles human relations, creates loneliness and deteriorates human relations

# Responsible Machines and Unexplainable Decisions

# Attribute Moral Responsibility

- Formal organizations have learned principles of accountability through responsibility attribution
- If AI is given decision agency – how to attribute moral responsibility? Harms? Benefits?
- Moral responsibility – if you have effects on the world, you are responsible for consequences
  - Aristotle's Nicomachean Ethics – actions have their origin in the agent – normative side: if you take action, you take responsibility
  - Aristotle adds: **you are responsible if you know what you are doing**



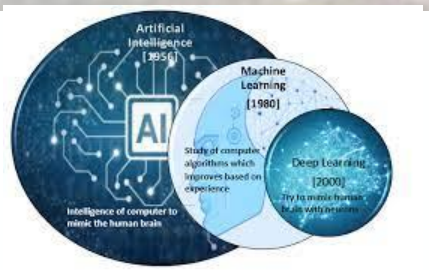
# AI Moral Attribution

- AI can make decisions/actions
  - Lack awareness and moral thought, lack consciousness
    - Is there capability of intention
  - In this light humans can delegate agency to machines but retain responsibility
    - Not different from structured organizations – managers retain responsibility
    - Not different from dogs and small children – caretakers are responsible
- This approach is challenging for AI agents

# AI Moral Attribution

- AI agents operate at high frequency
  - AI stock trading; self-driving vehicle
  - Insufficient time for human review
- Who is the "caretaker"?
  - AI developer? First user? Second user?
  - E.g., AI developed at university, then applied in health sector, then moved to military use
- Example – March 2018 self-driving Uber killed pedestrian
  - Programmer responsibility?
  - Car Manufacturer?
  - Uber?
  - Car user?
  - Pedestrian?
  - Arizona regulators?
- Often it is not one person or one component
  - AI algorithms interact with sensors, collects data, interacts with HW…
  - ML trained with data sets, collected in different ways, treated in different ways

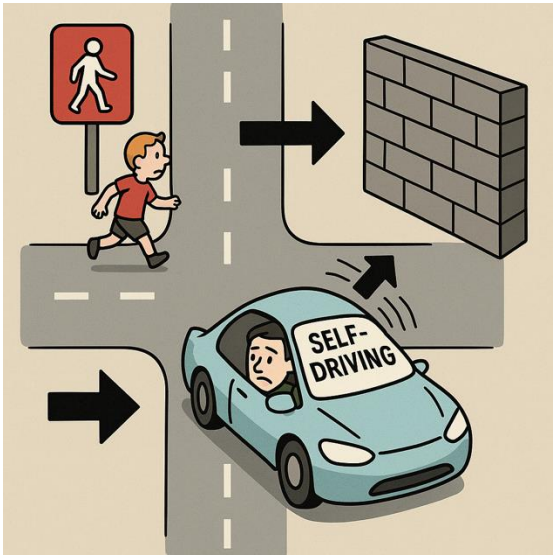
# Key Questions




- How to attribute/distribute responsibility?
- How do we determine needed knowledge?
  - Responsibility entails answerability and explainability
  - AI does not "know" what it is doing the way we intend
    - Lack of consciousness – it can record what done but no awareness to deliberate and reflect on actions [that's why children are not hold responsible]
    - If human responsible – how do they know?
- Statistical-based AI systems (Machine Learning) are problematic
  - What happens in the layers of network cannot be explained
  - Black box problem
    - Even developers may not know the AI system (e.g., libraries, successive development teams, programmers vs data providers)
    - Programmers may not know future uses of the algorithms (e.g., unintended uses)
- Lack of transparency/explainability also leads to lack of trust

# Moral Decision and Human Control

- Other challenges
  - Lack of "moral" standards in automated decision making
  - Tension between human and automated control
    - 2016 truck hijacked in Berlin
    - 2018/2019 Boeing 737 MAX
    - Existing guidelines (SAE J3016) analyzes complementarity, not conflict
      - SAE J3016 prominent technical document that defines levels of driving automation

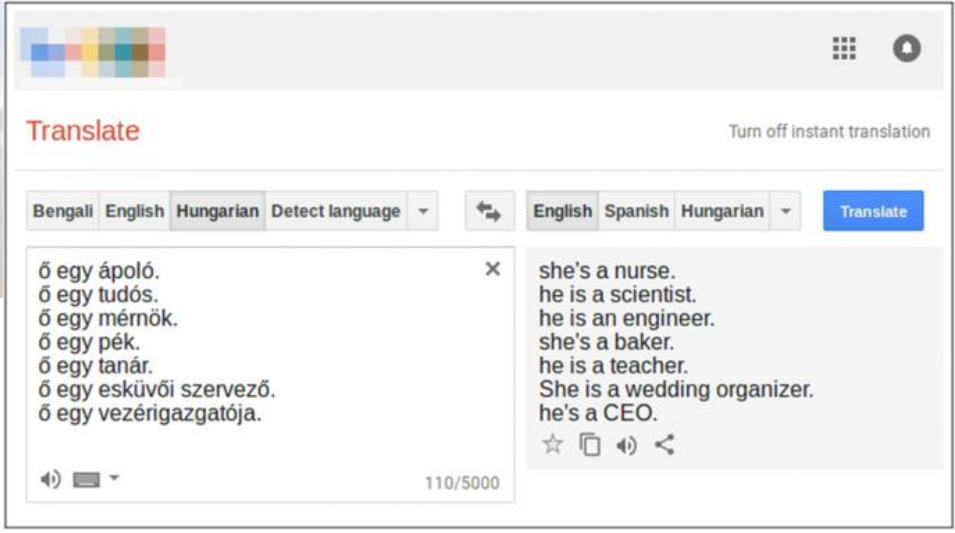


# Biases and the Meaning of Life

# Bias

- When an AI makes—or recommends—decisions, bias may arise
  - The decisions may be unjust or unfair to particular individuals or groups
  - Frequently associated to ML
- Biases exist in modern society – AI may amplify them
- Often is **unintended** by developers
  - Do not understand system, do not reflect on own biases, unintended uses of technology
  - Profound consequences – not getting a job or a loan, entire communities labeled as high risk
  - COMPAS – prediction of re-offense of convicted criminals
    - Used in Florida to recommend parole
    - False positives – predominantly black
    - False negative - predominantly white
  - PREDPOL – predictive policing tool – allocate police to areas of town
    - Prevailing allocation of police force in poor neighborhoods
    - Break down in trust between law enforcement and public

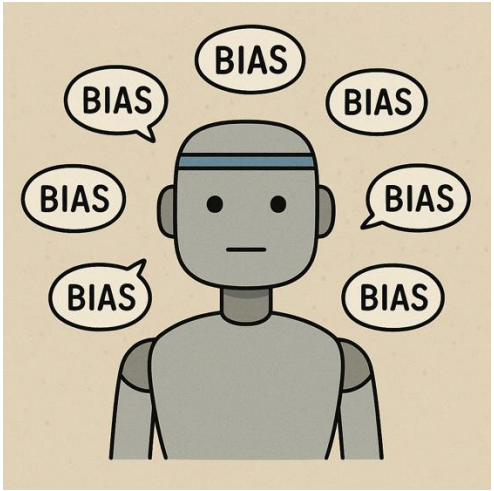
# Bias




- Bias introduced in all stages of machine learning development
- Examples
  - Data not representative
    - Collect data about American White Males but used for predictions on the entire population
  - Difference among countries
    - ImageNet predominantly based on US images
    - Cultural Bias
  - Insufficient data – e.g., there are not "that many" murders
  - Biases embedded in society – medicine uses is not equitable
    - Primarily used by male patients from middle/upper class
  - NLP gender bias of existing training corpora, e.g., analogies systems (Reddit data)
    - Man is to king as woman is to X (X=queen)
    - Man is to programmer as woman is to X (X=homemaker)

# Bias

- Are unbiased approaches possible?
  - Could we compromise precision of prediction for unbiased?
  - Bias permeates all aspects of society – impossible to fully remove
    - Datasets are abstractions of society
    - Effort to mitigate
  - Discrimination is sometimes required
    - CV screener to recommend for a position
    - Spotify recommender tries to follow your preferences
      - Is this biased against lesser known artists?
  - Debate about whether AI should mirror the real world or actively improve it
    - Google search favors male over female science professors in searches
      - Should Google prioritize images of female science professors to adjust?



# Future of Work and Meaning of Life

- AI positioned to transform society – work and life
- Will AI destroy jobs?
  - What kind of jobs? Blue Collar jobs? Advanced cognitive systems may impact all kind of jobs (47% of jobs requiring high school diploma)
  - Wider gap between rich and poor? Educated and uneducated? Technologically advanced countries vs developing countries?
- Utopian predictions of leisure societies have not been achieved
  - Automation introduced since 19th century
  - No signs of liberation and emancipation
  - Hierarchical structure of society remains rigid (some have benefited enormously, others not at all)
  - Work is not necessarily punishment
    - Opportunity for social connection, sense of belonging, exercise responsibility
    - Do not delegate to AI work that is creative and enjoyable

# Thank you