CSC520 - Artificial Intelligence

Lecture 26

Dr. Scott N. Gerard

North Carolina State University

Apr 22, 2025

Agenda

- Ethics of Software Professionals
- Benefits of AI
- Risks of AI
- IEEE principles of Ethically Aligned Design

Joint ACM/IEEE-CS Software Engineering Code

https://www.acm.org/code-of-ethics/software-engineering-code/

1. **PUBLIC**: SEs shall act consistently with the public interest.
2. **CLIENT AND EMPLOYER**: SEs shall act . . . in the best interests of their client and employer consistent with the public interest.
3. **PRODUCT**: SEs shall ensure that their products and related modifications meet the highest professional standards possible.
4. **JUDGMENT**: SEs shall maintain integrity and independence in their professional judgment.
5. **MANAGEMENT**: SE managers and leaders shall subscribe to and promote an ethical approach to the management of software development and maintenance.
6. **PROFESSION**: SEs shall advance the integrity and reputation of the profession consistent with the public interest.
7. **COLLEAGUES**: SEs shall be fair and supportive of their colleagues.
8. **SELF**: SEs shall participate in lifelong learning regarding the practice of their profession and shall promote an ethical approach to the practice of the profession.

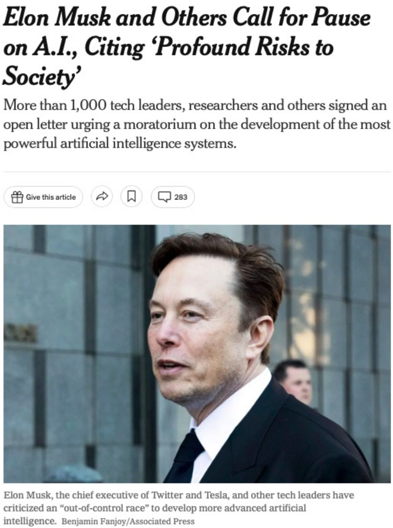
Benefits of AI

- Improved healthcare
  - Accurate cancer diagnosis
  - Robot/AI assisted surgeries
- Crop management and food production
  - AI/ML improve crop yield prediction using sensor data and visual analytics data from drones
  - Smart tractors, agribots and robots for agricultural operations
- Assist people with disabilities in seeing, hearing and mobility
  - Text-to-speech conversion for blind
  - Smart speakers for voice interaction, playing music, controlling home devices, etc.
- Increase productivity through factory automation
  - Robots/AI automate repetitive tasks
- Reduce risks to human lives by performing dangerous tasks
  - Robots/AI in mining and utilities maintenance
  - Robots/AI in nuclear power plant operations

Risks

- **Reputational Risks**
- **Financial Risks**
- **Legal Risks**

Risks of AI



Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'

More than 1,000 tech leaders, researchers and others signed an open letter urging a moratorium on the development of the most powerful artificial intelligence systems.

Elon Musk, the chief executive of Twitter and Tesla, and other tech leaders have criticized an "out-of-control race" to develop more advanced artificial intelligence. Benjamin Fanjoy/Associated Press

Source: New York Times, March 29, 2023

Risks of AI

- “Pause Giant AI Experiments: An Open Letter”, published on March 22, 2023, signed by over 1000 AI leaders
- Questions raised in the letter
  - Should we let machines flood our information channels with propaganda and untruth?
  - Should we automate away all the jobs, including the fulfilling ones?
  - Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us?
  - Should we risk loss of control of our civilization?
- Letter calls on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4
- Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.

Risks of AI

- “Pause Giant AI Experiments: An Open Letter”, published on March 22, 2023, signed by over 1000 AI leaders
- Questions raised in the letter
  - Should we let machines flood our information channels with propaganda and untruth?
  - Should we automate away all the jobs, including the fulfilling ones?
  - Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us?
  - Should we risk loss of control of our civilization?
- Letter calls on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4
- Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.
- **There was no pause**

Risks of AI

- Autonomous lethal weapons
  - Should we delegate the decision to kill humans to a machine?
  - E.g., Israel's Harop missile searches for, decides to engage, and engages targets on its own without human intervention
- Mass unemployment
  - Machines can replace humans even for complex tasks
- Violation of privacy
  - Communications (phone, internet) data, medical records, etc. may be collected
- Bias and discrimination
  - ML models learn and perpetuate societal bias
- AI in wrong hands
  - E.g., bad actors can misuse autonomous weapons, carry out cyberattacks, etc.

Reputational Damage [1]

- Google Image Misclassification: Google's AI misclassified images of Black people as gorillas, leading to removal of the feature. (2015)
- Microsoft's Tay Tweets: The Tay chatbot learned and posted inappropriate content on Twitter due to user manipulation, forcing Microsoft to shut it down within 24 hours. (2016)
- Google Hallucination: Google's Bard chatbot wrongly claimed the James Webb Space Telescope was used to take the first pictures of exoplanets (2023)
- Air Canada's Chatbot Error: Air Canada's AI tool gave incorrect advice for securing bereavement fare tickets, leading to legal action and partial refunds of $812 CAD and reputational damage. (2024)

[1] https://www.enkryptai.com/blog/how-to-prevent-ai-hallucinations

Avianca Airlines

- Roberto Mata sued Avianca saying he was injured when a metal serving cart struck his knee
- Lawyer Steven A. Schwartz (Levidow, Levidow Oberman) used ChatGPT to create a legal brief
  - Included more than half a dozen relevant court decisions
  - Submitted brief to the court
  - Neither judge nor opposing council could find NON-EXISTING citations
- Schwartz LoDuca fined $5,000 for citing made-up cases. "The Court concludes that a penalty of $5,000 . . . is sufficient to advance the goals of specific and general deterrence"
- Judge in Southern District of New York said lawyers
  - "abandoned their responsibilities" by submitting non-existent judicial opinions
  - "continued to stand by the fake opinions after judicial orders called their existence into question."

UK Horizon Post Office

- In 1999, Horizon computer system introduced to UK Post Office
  - Between 2000 and 2014, UK Post Office prosecuted 736 staff resulting in many convictions for false accounting and theft
  - Many died from related health conditions under a dark and stressful cloud of civil criminal proceedings
  - Many were financially ruined
- Horizon's broken algorithm was treated as indisputable truth
- In 2021, 12 years after problems were first publicly published,
  - 39 subpostmasters' convictions were quashed at the Court of Appeal
  - Post Office "should not have prosecuted them in the first place"
  - Post Office's conduct was "an affront to the conscience of the court"
  - Post Office reviewed Horizon, and lied about its findings
- Horizon was a deeply flawed computer system

Regulations

- US
  - **FERPA** (1974): protects the privacy of student education records
  - **HIPAA** (1995, 2003, 2006): governs the use and disclosure of Protected Health Information (PHI)
  - No GenAI laws or policies. Several states are investigating.
- EU
  - **GDPR** (2018): protect the rights and freedoms of individuals, ensuring control over personal data and harms in processing
  - **EU Data Act** (2023): risks related to data access, usage, and sharing across various sectors
  - **EU AI Act** (2024): categorizing risk into "unacceptable," "high," "limited," and "minimal"
- China: regulations about algorithmic recommendations, deep synthesis, and generative AI
- India: No laws, but policies, guidelines, and sector-specific regulations

Ethics

- All of us have a moral obligation to promote the positive aspects of AI and avoid or mitigate its negative aspects
- Academia, industry, and governments around the world are actively working on addressing the ethical concerns
- Several ethical guidelines and principles have been proposed

IEEE Ethically Aligned Design

https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

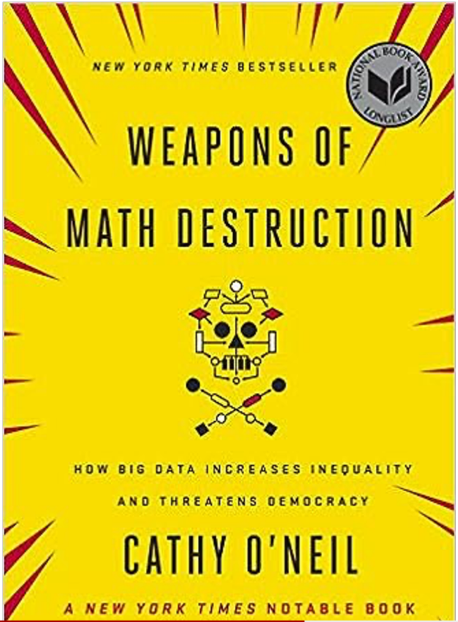
Autonomous and Intelligent Systems (A/IS) is broader than AI

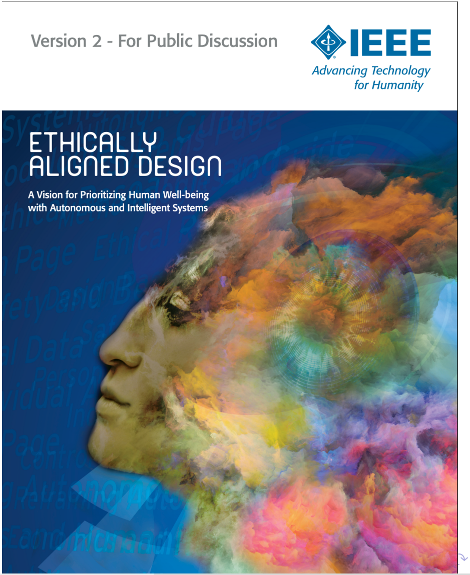
- **Human Rights**: How can we ensure that A/IS do not infringe upon human rights?
- **Prioritizing Well-being**: Traditional metrics of prosperity do not take into account the full effect of A/IS technologies on human well-being
- **Accountability**: How can we assure that designers, manufacturers, owners, and operators of A/IS are responsible and accountable?
- **Transparency**: How can we ensure that A/IS are transparent?
- **A/IS Technology Misuse and Awareness of It**: How can we extend the benefits and minimize the risks of A/IS technology being misused?

Ethics Survey – Optional

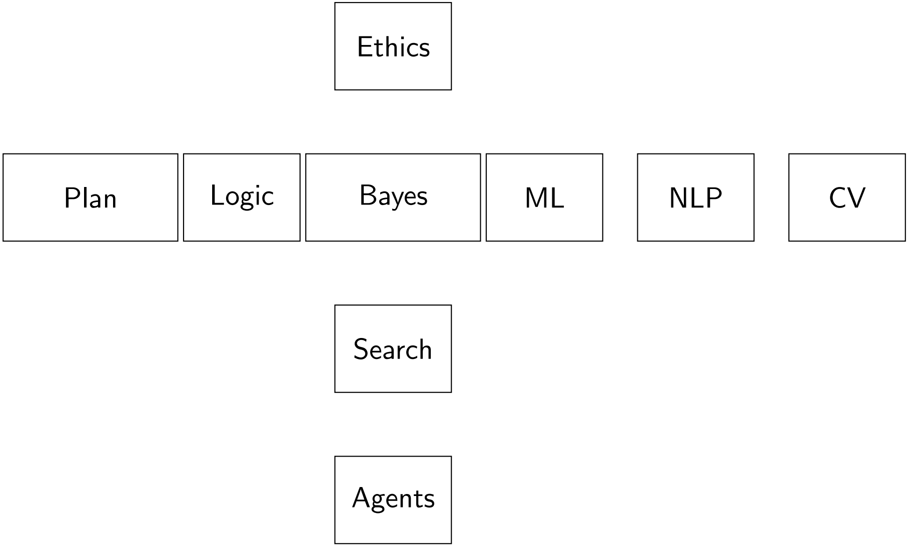
- Niraj Sitaula (nsitaula@syr.edu), a PhD candidate in the School of Information Studies at Syracuse University.
    - His doctoral adviser, Dr. Jennifer Stromer-Galley (jstromer@syr.edu)
- Conducting a survey of Indian international students and American students studying AI and data science in the US to examine differences in perspectives around ethics and trust in AI among different student populations
- 15 minutes to complete.
- Students who complete the survey will be entered into a lottery to receive 1 of 50 $20 Amazon gift cards

Valuable Reading


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK


Version 2 - For Public Discussion
IEEE
Advancing Technology for Humanity
ETHICALLY ALIGNED DESIGN
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

We’ve Come a Long Way



Ethics

Plan

Logic

Bayes

ML

NLP

CV

Search

Agents

AI and AI-Adjacent Careers

- AI Engineer / Data Scientist
- AI Academic / Researcher / Educator

AI and AI-Adjacent Careers

- AI Engineer / Data Scientist
- AI Academic / Researcher / Educator

- Data Engineer
- Visualization / User Interface
- AI Ethicist
- Manager / Product Manager / Project Manager for AI team

AI and AI-Adjacent Careers

- AI Engineer / Data Scientist
- AI Academic / Researcher / Educator

- Data Engineer
- Visualization / User Interface
- AI Ethicist
- Manager / Product Manager / Project Manager for AI team

- You already know more AI than most of the planet

AI and AI-Adjacent Careers

- AI Engineer / Data Scientist
- AI Academic / Researcher / Educator

- Data Engineer
- Visualization / User Interface
- AI Ethicist
- Manager / Product Manager / Project Manager for AI team

- You already know more AI than most of the planet

- You must continually study hard to stay current

Final Exam

- Date/time: 4/24/2023 from 8:30 AM – 11:00 AM = 2.5 hours
- Location: Usual classroom
- Comprehensive exam
  - 10 Questions
  - All topics from the beginning are fair game
  - Questions similar to homeworks, Midterm I, Midterm II
  - Conceptual questions
  - Problem solving
- Bring a simple calculator to the exam