CYB 4203/6203

Secure and Trustworthy AI

Presentation 4: Human Rights and Human-AI Collaboration

Wednesday, February 4, 2026

Topics: 2.2, 2.3, 2.4

Recap – Last Week

Key Themes: 2.2 – Core AI Values (part I)

- **Fairness** types: distributive, procedural, interactional, personal vs. collective
- Fairness challenges: context-dependence, provable impossibility, algorithmic (definition, measurement, optimization), data (world we have vs. want)
- Fairness re: core machine learning tasks – (1) classification, (2) regression & prediction, (3) ranking & recommendation, (4) clustering & segmentation, (5) generation, (6) detection & recognition, (7) optimization & resource allocation

Recap – Last Week

Key Themes: 2.2 – Core AI Values (part I)

- **Transparency**: Model, data, decision, trade-offs, stakeholder-appropriateness
- **Accountability**: Ownership, governance, redress, liability, challenges

Today’s Agenda

- 2.2 Core AI values:
    - Privacy
    - Autonomy
    - Safety
    - Sustainability
- 2.3 AI and human rights
- 2.4 Human-AI collaboration

Topic 2.2: Core AI Values

Fairness – Transparency – Accountability

Privacy – Autonomy – Safety – Sustainability*

*not an exhaustive list

Core AI Value - Privacy

AI needs vast datasets – often collected without consent

- Privacy threats: deepfake ID fraud, automated phishing, inference of sensitive info from seemingly unrelated data, etc…
- Users sharing data with chatbots may not read TOS
    - Stanford study: privacy risks of AI chatbot conversations
- Fragmented regulatory landscape – GDPR, CCPA, EU AI Act
    - 2025: 1262 AI-related laws proposed in the U.S. (NCSL database)
- Machine unlearning – technical process of removing the influence of specific training data from a ML model without retraining from scratch

Core AI Value - Autonomy

“You can choose a ready guide in some celestial voice

If you choose not to decide, you still have made a choice

You can choose from phantom fears and kindness that can kill…”

(what’s the next lyric?)

Core AI Value - Autonomy

"You can choose a ready guide in some celestial voice
If you choose not to decide, you still have made a choice
You can choose from phantom fears and kindness that can kill
I will choose a path that's clear: I will choose Freewill."

- Neil Peart (1952-2007), Freewill by Rush

Core AI Value - Autonomy

”… a person’s effective capacity to act on the basis of beliefs, values, motivations, and reasons that are in some relevant sense their own.”

- Carina Prunkl, 2024

Prunkl discusses several dimensions of autonomy:

- Authenticity (internal): Are my beliefs truly ‘mine’?
  - Example: adaptive preference formation from limited presented options
- Agency (external): Can I act on my beliefs to control my life?
  - Competency (skill), freedom (legal), opportunity (social access)
- Discussion: How is AI impacting internal and external autonomy?

Core AI Value - Safety

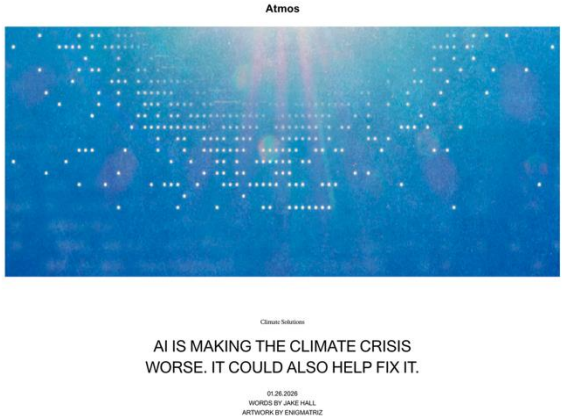
Is it possible to make intelligence ‘safe’?

Core challenges: Alignment, Control, Interpretability, Robustness

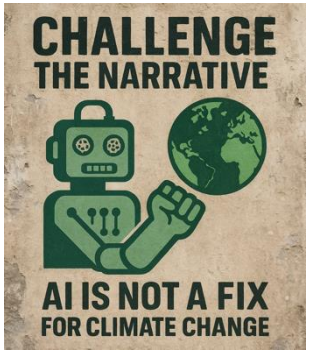
- February 2026 International AI Safety Report highlights:
  - Gen-AI capabilities continue to improve, esp. math, coding, autonomy
  - Fueled by both pre-training and post-training
  - Rapid (faster than PC) but unevenly-distributed adoption
  - Heightened concerns re: bio-weapons development, cyberattacks
- Future of Life Institute Winter 2025 Safety Index Report
  - AI experts rate leading AI companies on key safety/security domains

Core AI Value – Sustainability (click the images!)

IBM video
(pro)



Atmos article
(neutral?)



Medium article
(con)

Topic 2.3: AI and Human Rights

Cross-cultural, Legal, and Environmental Perspectives

AI and Human Rights: Cultural & Legal perspectives

A mix of good and bad news:

- AI Bias & Representation
  - Amazon Hiring Discrimination - Google Vision Racism
- Creative Labor & Displacement
  - SAG-AFTRA WGA – Anderson et al. v. Stability AI – Bartz v. Anthropic
- Misinformation & Cultural Trust
  - Political Deepfakes – AI Slop Journalism Spread
- Invisible Labor & Exploitation
  - OpenAI $2/hr Content Filtration – Empire of AI (Karen Hao)
- Language & Knowledge Erasure
  - Language Representation inequality – AI Preservation of Endangered Languages

AI and Human Rights: Environmental perspectives

Key concerns: Energy / water use. Mining / mfg. Unequal distribution.

- MIT article: Generative AI's environmental impact
- GenAI clusters can consume 7-8x energy of typical workloads
  - GPT-3 training: estimated 1,287 MWh (~120 homes for a year)
- 2026 projection: 1,050 TWh globally – 5th place (b/t Japan & Russia)
- 2030 projection: ~34 M metric tons CO^2 (7.5M cars), ~1B m^3 water (1 km^3)
  - Intuition: Earth has ~1.5B cars & ~41M km^3 accessible fresh water
- Mining / manufacturing paradox:
  - AI hardware → increased need for rare earth minerals
  - AI-enhanced exploration, discovery, mining, processing, mfg.

Topic 2.4: Human-AI Collaboration

Designing systems that augment humans
(rather than replacing them)

Human-AI Collaboration

Keeping it light here – but there are many more examples out there.

- Centaurs & Cyborgs on the Jagged Frontier – Ethan Mollick
- Real-world examples of Human-AI Collaboration (unabashedly pro-AI perspective) – SmythOS
- Human-AI Teaming in Healthcare – Nature Artificial Intelligence (NPJ)

Assignments for this week

Assignment 3: Core AI Values Case Study

- Similar to Assignment 2, but you choose pro / con angle
- I'll release full details tomorrow at 12:30 pm

Assigned Reading:

Hendrycks, Chapter 1: Overview of Catastrophic AI Risks

NEXT WEEK:

Potential Harms & Misuse

- 3.1 Unintended harms (briefly)
- 3.2 Intentional misuse (more deeply)
- 3.3 The alignment problem (omg)
- 3.4 Responsible AI innovation practices & obligations

See you next time :)

This presentation was drafted and refined using Claude 4.5 Sonnet and Claude Code.

All content was reviewed and approved by Dallas Elleman.