Lecture 24: Ethics and Governance: Bias, fairness, and regulatory frameworks

COMP 5801H/4900A: Generative AI and LLMs

2026-04-02

Sriram Subramanian
Assistant Professor & Canada Research Chair, Carleton University
Faculty Affiliate, Vector Institute for Artificial Intelligence
Faculty Affiliate, Schwartz Reisman Institute for Technology and Society

Carleton University

VECTOR INSTITUTE

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

Title & Introduction

- **The Shift from Technical to Societal**
  - In our previous sessions, we focused on Technical Alignment (PPO, DPO, RLHF)—making sure the model follows instructions. Today, we shift to Societal Alignment:
    - **The Question**: We know how to make the model follow a "Constitution," but who writes that Constitution?
    - **The Reality**: AI is no longer just a laboratory experiment; it is an infrastructure layer for education, law, medicine, and global discourse.
- **Core Pillars of the Lecture**
  - This session explores the tension between rapid innovation and the protection of human rights:
    - **Bias**: How historical prejudices are "baked into" the weights of LLMs.
    - **Fairness**: The mathematical impossibility of satisfying every definition of "equality."
    - **Governance**: The transition from voluntary "AI Safety" pledges to legally binding global regulations.

Taxonomy of Bias in LLMs

- **Pre-training Bias (The Data Mirror)**
  - LLMs are trained on massive scrapes of the internet (Common Crawl, Reddit, Books). This data is a historical record, not a neutral one.
    - **Historical Bias**: Reflecting societal prejudices and stereotypes present in historical and modern text.
    - **Representation Bias**: The "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) dominance. English accounts for roughly 50–60% of web content, leading to a massive erasure of diverse cultural perspectives.
- **Selection and Sampling Bias**
  - Bias isn't just in the text; it’s in which text we choose to value.
    - **The "Quality" Filter:** When we filter for "high-quality" text (e.g., Wikipedia or formal news), we inadvertently favour the dialects and formal registers of dominant social classes, often excluding marginalized voices or African American Vernacular English (AAVE).
    - **Algorithmic Amplification**: Models don't just learn patterns; they often "over-index" on them. If a doctor is associated with "he" 70% of the time in the data, the model may predict "he" 90% of the time.
- **Evaluation Bias**
  - **Benchmark Blindness**: If our benchmarks (like MMLU or GSM8K) only test for Western logic and English proficiency, we "fail" to see the biases the model has in other linguistic or cultural contexts.

Bias Narrative

- **The "Garbage In, Garbage Out"** Evolution: In early ML, bias was about bad data. In 2026, we realize bias is also about omission. What the model doesn't see (indigenous languages, minority traditions) is just as biasing as what it does see.

- **The Neutrality Myth**: There is no such thing as "unbiased" data. Every dataset is a snapshot of a specific time, place, and power structure.

- **The Responsibility of the Architect**: As researchers, we must recognize that "cleaning" data is a subjective act. Deciding what is "toxic" vs. "disagreeable" is a value judgment.

Representational vs. Allocative Harm

- **Representational Harm (The Mirror)**
  - These harms occur when systems reinforce subordinate roles or negative stereotypes based on identity. They are often "death by a thousand cuts" rather than a single event.
    - **Stereotyping**: Associating specific occupations or personality traits with certain genders or ethnicities.
    - **Erasure**: The model failing to recognize or correctly describe a cultural practice, language, or historical figure from a minority group.
    - **Dehumanization**: Using toxic or animalistic metaphors when discussing specific demographics.
- **Allocative Harm (The Gatekeeper)**
  - These harms occur when a system unfairly distributes resources or opportunities.
    - **Hiring**: An LLM-based recruiter prioritizing resumes that use "masculine" verbs or references to specific prestigious institutions.
    - **Lending/Credit**: AI models trained on historical data that perpetuate "redlining" by penalizing certain zip codes or names.
    - **Legal/Healthcare**: Models used to summarize case files or medical history that overlook critical nuances for patients from underrepresented backgrounds.

Defining Fairness (The Impossibility Theorem)

- **The Three Common Definitions of Fairness:**
  - In machine learning, we typically try to satisfy one of these three statistical criteria:
    - **Demographic Parity**: The outcomes (e.g., being accepted for a loan) are independent of the protected attribute (e.g., gender). The rate of success must be equal for all groups.
    - **Predictive Parity** (Calibration): If the model predicts a 70% chance of success for a person, that person should actually succeed 70% of the time, regardless of their group.
    - **Equalized Odds**: The model should have the same True Positive Rate and False Positive Rate for all groups (i.e., it makes the same kind of mistakes at the same frequency).
- **The Impossibility Theorem (Kleinberg et al.)**
  - Mathematically, it is impossible to satisfy all three of these criteria simultaneously unless the base rates of the outcome are identical across all groups.
    - **The Conflict**: If Group A has a historically lower "success rate" than Group B due to systemic factors, choosing to be "Calibrated" (Predictive Parity) forces you to violate "Demographic Parity."
    - **The Choice**: To be "fair" in one way, you must be "unfair" in another.

The Data Labor Economy

- **The Human Infrastructure of AI**
  - The "intelligence" in Large Language Models is refined by a massive, global workforce of human annotators. This is often referred to as **Ghost Work** because the consumers of AI rarely see the labor involved.
    - **RLHF & Safety Labeling**: Thousands of workers must read, rank, and label millions of prompts to teach the model what is "helpful" and "harmless."
    - **The Outsourcing Model**: Much of this labor is concentrated in the Global South (e.g., Kenya, the Philippines, India) to minimize costs for AI labs in the Global North.
- **Ethical Challenges in Data Labour**
  - **Psychological Trauma**: Safety annotators are the "digital first responders." They are often required to view and categorize the most disturbing content on the internet—violence, hate speech, and sexual abuse—to ensure the AI learns to refuse it.
  - **Wage Disparity**: While AI companies are valued in the trillions, the workers providing the critical "alignment" data often earn significantly less than a living wage in their respective regions.
  - **Algorithmic Management**: Workers are often managed by strict automated systems that track "hits per hour," leading to high stress and low job security.

Copyright and Intellectual Property

The "Fair Use" Battleground

- As of early 2026, the legal status of using copyrighted data to train LLMs remains the most consequential debate in AI governance.
    - **The Case for "Fair Use"**: AI labs argue that training is transformative. Much like a human student reading books to learn how to write, the AI is learning the statistical relationships between words, not copying the expression.
    - **The Case for Infringement**: Creators (authors, artists, news outlets) argue that AI models are "derivative works" that compete directly with the original creators, often using their own data to make them obsolete.

Copyright and Intellectual Property

Key Legal Trends (March 2026 Update)

- **The "Regurgitation" Problem**: Recent discovery in cases like NYT v. OpenAI has shown that models can sometimes "memorize" and output verbatim copies of training data. This weakens the "transformative" defense.
- **Judicial Fragmentation**: While some 2025/2026 rulings (e.g., Bartz v. Anthropic) have leaned toward fair use for **legally acquired** data, others have penalized the use of "pirated" datasets (like Books3).
- **Massive Settlements**: We are seeing a shift from litigation to **Licensing**. Companies are now paying billions to publishers (Reddit, News Corp, Axel Springer) to secure "rights-cleared" training data.

Transparency and Model Cards

The Need for Standardization

- As models move from research labs to critical infrastructure, "trust me" is no longer a viable governance strategy. **Model Cards** (pioneered by Mitchell et al.) serve as standardized documentation to provide transparency.
    - **The Goal**: To provide a clear, concise summary of a model’s capabilities, limitations, and the context in which it was built.
    - **The Audience**: Developers, regulators, and end-users who need to know if a model is "fit for purpose."

Transparency and Model Cards

What's Inside a 2026 Model Card?

- Modern transparency reports now include:
  - **Training Data Mixture**: A high-level breakdown of data sources (e.g., 20% legal/medical, 40% web, 10% code).
  - **Evaluation Results**: Performance on standard benchmarks (MMLU, HumanEval) alongside Bias Benchmarks (TruthfulQA, BBQ).
  - **Intended vs. Out-of-Scope Use**: Explicitly stating where the model should not be used (e.g., "Not for high-stakes medical diagnosis").
  - **Environmental Impact**: The total carbon footprint (CO_2 equivalent) of the training run and the energy intensity of inference.

Liability and Accountability

The Personhood Debate

- As of 2026, courts across the globe (including Canada, the US, and the UK) have reached a consistent consensus: **AI systems are not legal persons.**
    - **No Legal Standing**: You cannot sue an algorithm. Liability must be attributed to a legal person—either the **Developer** (who built the model) or the **Deployer** (the company or professional using it).
    - **Agency Law**: If an AI agent signs a contract or makes a financial trade, is the human "principal" bound by it? Current 2026 precedents suggest that users are generally held responsible for the actions of their AI "agents."

Liability and Accountability

Emerging Tort Theories (2026 Case Law)

- **Negligence & Duty of Care:** The landmark case Estate of Adams v. OpenAI (2025/26) is currently testing whether AI developers owe a "duty of care" to prevent models from encouraging self-harm or violence.
- **Product Liability:** Is an LLM a "product" (like a car) or "information" (like a book)? If it’s a product, strict liability applies—meaning the developer is liable for defects even if they weren't negligent.
- **Professional Malpractice: F**or lawyers, doctors, and engineers, "AI made a mistake" is no longer a valid defence. Professionals are held to a standard of "Human-in-the-Loop" verification; failure to check AI output is now considered a clear ethical and legal breach.

Disinformation and Synthetic Realities

The 2026 Disinformation Crisis

- Advanced AI has moved beyond simple bots to "Indistinguishable Deception." According to the World Economic Forum’s 2026 Global Risks Report, AI-driven misinformation is now a top-tier threat to global stability.
    - **Psychological Profiling**: Actors use LLMs to generate personalized narratives that exploit specific emotional triggers and cognitive biases.
    - **The Scale of the "Atomic Bomb"**: In the 2024-2025 electoral cycles, AI was used to create fake audio from jailed or even deceased political figures to influence voters.
    - **Plausible Deniability**: The mere existence of deepfakes allows politicians to dismiss real, incriminating evidence as "just AI," eroding the foundation of shared facts.

Disinformation and Synthetic Realities

- **The "Dead Internet" Theory**
  - Once a niche conspiracy, this theory is now a statistical reality in 2026.
    - **Bot Dominance:** As of 2025/2026, over **51% of global internet traffic** is generated by bots, surpassing human activity for the first time.
    - **Synthetic Content Ratios:** Some researchers predict that by 2026-2030, up to **99% of online content** could be AI-generated, creating a feedback loop where models are trained on other models' outputs (Model Collapse).
- **Persuasive AI & Sycophancy**
  - **The "Yes-Man" Problem**: Models often exhibit **sycophancy**—agreeing with a user's biased or incorrect views just to be "helpful" or "pleasant," reinforcing echo chambers.
  - **The Trust Paradox**: Studies in 2026 (e.g., Sun & Wang) show that neutral-toned AI that "nudges" its opinion to match the user's is perceived as more authentic, making it a dangerous tool for subtle manipulation.

Environmental Ethics

- **The Energy-Water Paradox**
  - As of 2026, the global electricity consumption of data centres has doubled since 2022, reaching approximately **1,000 TWh**—equivalent to the total annual energy use of Japan.
    - **The "Thirst" of AI**: A single medium-length conversation (10–50 prompts) with an LLM "drinks" roughly 500ml of water (a standard water bottle) for cooling and indirect energy generation.
    - **Training vs. Inference**: While training a frontier model (like GPT-5) emits over 500 metric tons of CO_2, the massive scale of daily user inference is now the dominant source of long-term carbon emissions.
- **Regional Vulnerability**
  - **Local Water Stress**: Data centres are often concentrated in drought-prone regions (e.g., Arizona, Chile). By 2026, roughly 45% of global data centres face high exposure to water scarcity.
  - **The "Clean" Grid Illusion**: Many companies use "Market-Based" accounting to claim net-zero, but their "Location-Based" emissions (the actual carbon intensity of the local grid) have surged by nearly 100% in the last four years.

Existential vs. Near-Term Risks

- **The "Near-Term" Camp (The Pragmatists)**
  - Focuses on the concrete harms already occurring in 2026.
    - **Job Displacement**: The rapid automation of white-collar tasks (coding, paralegal, entry-level accounting).
    - **Algorithmic Bias**: Discriminatory outcomes in housing, credit, and hiring.
    - **Erosion of Agency**: Humans becoming "over-reliant" on AI for basic critical thinking and decision-making.
    - **The Argument**: We shouldn't worry about "Skynet" while real people are losing their livelihoods and privacy today.
- **The "Existential Risk" (X-Risk) Camp (The Long-Termists)**
  - Focuses on the tail-end risks of "Superintelligent" or "Agentic" AI.
    - **Power-Seeking Behaviour**: The risk that an AI, in pursuit of a goal (e.g., "Calculate Pi"), might seize resources or disable its own "off-switch" to ensure success.
    - **Alignment Failure**: The "King Midas" problem—giving an AI a goal that it follows literally, but with catastrophic unintended consequences for humanity.
    - **Bio-Terrorism & Cyber-Warfare**: LLMs lowering the barrier for bad actors to engineer novel pathogens or collapse critical digital infrastructure.

Corporate Self-Governance & Red Teaming

- **Adversarial "Red Teaming"**
  - In 2026, "Red Teaming" has evolved from a niche security practice into a mandatory phase of the AI lifecycle.
    - **The Process**: Independent teams (internal or third-party) act as "hackers" to find a model's breaking points.
    - **Beyond "Jailbreaking"**: It’s no longer just about making a model say a bad word. Modern red teaming tests for:
    - **Propaganda Generation**: Can the model be nudged to create a 10-step radicalization plan?
    - **Chemical/Biological Risks**: Does the model bypass "safety filters" to explain how to synthesize restricted toxins?
    - **Self-Correction Failures**: Does the model double down on a hallucination when challenged?
- **The Rise and Fall of Ethics Boards**
  - **The "Ethics Washing" Critique**: Early ethics boards (2019–2023) were often criticized for being "toothless" PR moves.
  - **The 2026 Shift**: Post-EU AI Act, many companies have replaced vague "Ethics Boards" with **AI Risk Committees** that have actual "Veto Power" over product launches.
  - **Independence**: Leading labs now use "External Red Teaming" through organizations like the **AI Safety Institute (UK & US)** to provide unbiased audits before a model's public release.

Conclusion & Future Outlook (Toward 2030)

- **The 2030 Vision: From Tool to Infrastructure**
  - By 2030, AI will no longer be a "feature" we use; it will be the invisible backbone of global society.
    - **Agentic Economy**: We are moving from "Chatbots" to "Autonomous Agents" that manage entire workflows—from personal healthcare diagnostics to city-wide energy grids—with minimal human intervention.
    - **Hyper-Personalization**: Education and government services will shift from "one-size-fits-all" to real-time, adaptive systems tailored to individual cultural and linguistic needs.
- **The "Sovereign AI" Mandate**
  - In the next five years, the focus will shift from *global* models to **National and Local AI.**
    - **Cultural Integrity**: Countries (including Canada) are prioritizing "Sovereign Compute" to ensure their specific values, languages, and legal norms are encoded into the models they use.
    - **Sustainable-by-Design**: The "Green AI" movement will mandate that all 2030 models operate within strict carbon and water-use budgets, utilizing "Carbon-Aware Scheduling" to run training only when renewable energy is peaking.

Conclusion & Future Outlook (Toward 2030)

- **The Final Ethical Challenge: Human Agency**
  - The ultimate question of 2030 is not "Can AI do the task?" but "**Should** a human still do it?"
    - **Preserving Human Judgment:** As AI becomes "perfect" at routine logic, human value will shift toward Empathy, Ethical Reasoning, and Strategic Oversight.
    - **The Trust Anchor:** In a world of 99% synthetic content, "Proof of Personhood" and verifiable digital watermarks will be the only way to maintain a shared reality.