AI, Ethics, and Society

01.10.2025

Recap - Last In-Class Assignment

A car manufacturer releases the first fully self-driving car system, powered by a proprietary neural network trained mostly on private roads.

There is no formal regulation at the time of release. Early deployment shows:

- Fewer overall accidents than human drivers
- But higher fatality rate when crashes occur
- Passengers are more likely to survive than others involved in crashes
- Accident data is used to continuously re-train the model

Key Ethical Problems

- Transparency & Explainability
- Responsibility & Accountability
- Safety
- Justice & Fairness
- (Privacy)

Your answers

Transparency & Explainability

- Black-box problem
  - The use of neural nets makes AV decisions hard to explain post-incident.
- Public trust erosion
  - If the system can’t explain its actions, it risks damaging societal trust in AI.
- Legal accountability
  - Lack of transparency obstructs investigation and blame assignment after crashes.
- Ethical concern
  - Victims and families are left without answers.
- Training and datasets
  - Possible bias in training data as a transparency issue.
- Societal impact
  - Lack of explainability affects society at large (e.g., public fear, regulatory deadlock).

Responsibility & Accountability

- Ambiguity of liability
  - Current regulation does not define who is responsible: the user, manufacturer, or system.
- Moral hazard
  - If manufacturers aren't held liable, they may lack incentive to improve safety.
- Passenger guilt
  - Even if passengers are not technically responsible, they may feel moral guilt.
- Compensation concerns
  - Lack of clear responsibility can leave victims without justice or fair compensation.
- Real-world example
  - Reference to the 2018 Uber crash: the backup driver was held liable, raising questions about legal scapegoating.
- Proposed solution
  - Several groups argued for strict manufacturer liability and pre-defined redress mechanisms.

Safety

- Crash severity vs. frequency
  - While total accidents may go down, fatality rate per accident increases.
- Vulnerability of non-users
  - Pedestrians, cyclists, and passengers of non-autonomous vehicles bear more risk.
- Ethics-by-design
  - Some groups called for safety to be embedded at the design stage, not post-deployment.
- Psychological safety
  - Passenger anxiety and the fear of losing control.
- Testing ethics
  - Legitimacy of using public roads as live testing grounds.
- Trade-offs
  - Dilemma of counting fatalities vs. non-fatal but serious injuries.

Justice & Fairness

- Risk distribution
  - Non-users are disproportionately harmed in crashes.
- Lack of consent
  - People not using AVs didn’t agree to be part of the “experiment”.
- Biased training
  - If systems are trained in clean, private-road environments, they may perform worse in disadvantaged or chaotic public settings.
- Survival asymmetry
  - AVs may prioritize protecting their own passengers, creating systemic unfairness.
- Fairness vs. utility
  - Should aim to minimize total harm or distribute harm fairly?

My answers

Transparency

Can system behavior and decision-making be understood, scrutinized, and explained?

Positive Aspects:

- System logs and sensor data can support analysis.
- Clear performance metrics are available (e.g., crash frequency).
- Some systems display what the AI "sees" to passengers (visual UI).

Transparency

Ethical Concerns:

- Black-box models: decisions cannot be meaningfully explained.
- No access to training data or test environments.
- Users cannot understand edge-case decisions (e.g., who gets prioritized in a crash).

Justice & Fairness

Are risks and benefits distributed fairly across all groups?

Positive Aspects:

- Potential for fewer overall accidents.
- No human bias (e.g., road rage, racial profiling).
- Benefits for people unable to drive (e.g., elderly, disabled).

Justice & Fairness

Ethical Concerns:

- Passengers have higher survival rates than pedestrians or cyclists.
- Poor performance in underrepresented environments.
- Access gap: wealthy benefit, others may bear the risk.

Responsibility & Accountability

Who is morally and legally responsible when harm occurs?

Positive Aspects:

- Systems log events in detail.
- Some companies accept liability under specific conditions.
- Removes human errors such as drunk driving.

Responsibility & Accountability

Ethical Concerns:

- Blurred responsibility (developer, manufacturer, user?).
- Updates may alter behavior without notice.
- Legal systems not yet adapted for autonomous agents.

Safety

Does the system minimize risk and harm?

Positive Aspects:

- Reduction in accident rates.
- Eliminates driver fatigue or distraction.
- Consistent legal compliance.

Safety

Ethical Concerns:

- Higher fatality rate when crashes do occur.
- Overfitting to controlled environments.
- Failure in edge cases can be catastrophic.

Privacy

How is data collected, stored, and used?

Positive Aspects:

- Can improve safety and traffic efficiency.
- Anonymized data can inform urban planning.
- Logs enable transparency and learning from errors.

Privacy

Ethical Concerns:

- Continuous tracking of users and bystanders.
- No meaningful consent from non-users.
- Risk of data reuse for commercial purposes.

ChatGPT’s answer to “Name one thing that is missing in this list”

Regulatory Vacuum and Precaution Principle

No exploration of the ethical obligation to delay deployment until robust policy frameworks are in place.

- Autonomous vehicles are not ethically deployable by default; doing so without regulation violates the precautionary principle (act with caution when consequences are uncertain and potentially irreversible).
- Governments share responsibility: “regulatory negligence” is ethically relevant.

Moving on…

Recap Reflection and Questions

- How should responsibility and liability be handled before deploying self-driving cars?

➢ Deployment without clear rules is ethically problematic. At minimum, there must be liability frameworks defining who is accountable for accidents (manufacturer, operator, insurer). Otherwise, victims risk being left without justice or compensation.

Recap Reflection and Questions

- Is it acceptable for self-driving cars to prioritize passenger safety over bystanders?

➢ No. Ethical design must not shift risk onto non-consenting bystanders. Principles of fairness and justice require that safety improvements benefit all road users, especially the most vulnerable.

Recap Reflection and Questions

- How do we apply ethical theories like utilitarianism and proportionality in practice?

➢ Utilitarianism considers both the number and severity of harms, but in practice it is difficult to weigh one death against several injuries. Proportionality asks if the benefits (fewer accidents, convenience) are justified compared to the risks and whether safer alternatives exist.

Recap Reflection and Questions

- When and how should AI systems be re-evaluated ethically if they continuously learn and adapt?

➢ Ethical assessment must be ongoing. Re-evaluation is needed after major updates, new deployments, or when evidence of harm emerges. Periodic audits and continuous monitoring are essential to ensure evolving systems remain aligned with ethical principles.

Moving on…

In-Class (or after class discussion)

Hypothetical example:

A university uses AI to flag students at risk of failing a course, based on attendance and grades.

- Ethics-by-Design
  - What values would need to be built into this system from the start? What technical or design decisions would need to reflect those values?
- Participatory Ethics
  - Who should be consulted or involved during the design phase? Why? Whose perspectives might otherwise be ignored?

What's one ethical concern that would be hard to solve through design alone?

Any suggestions?

Some of your answers

Ethics-by-Design (1/2)

Core values to embed:

- Fairness, transparency, accountability, privacy, accessibility.

Data and algorithm design:

- Audit training data for representation (gender, race, disability, etc.).
- Include fairness constraints to avoid systemic bias.
- Avoid irrelevant data collection; limit to what is strictly needed.

Ethics-by-Design (2/2)

System features:

- Explainability mechanisms so students understand why they were flagged.
- Human-in-the-loop oversight, especially for borderline cases.
- Clear, accessible communication of requirements and processes.
- Standardized procedures for special needs accommodations.
- Formal appeals processes with due process and legal safeguards.

Tensions:

- Transparency vs. privacy (how much explanation can be given without exposing sensitive criteria or data).

Participatory Ethics (1/2)

Students:

- Student Council (general representation).
- Student associations and sub-groups (specific needs, voices).
- Particularly vulnerable groups: disabled students, neurodivergent students, minority groups, students with jobs/family responsibilities.

Faculty & staff

- Teachers, department chairs, advisors, support services, IT/data officers (for compliance).

External expertise

- Ethicists, inclusion experts, accessibility specialists.

Participatory Ethics (2/2)

Processes suggested:

- Surveys and polls (anonymous, representative, recurring).
- Partitioning student groups and electing representatives.
- Regular re-evaluation and stakeholder interviews at project milestones.
- Structured communication channels for deliberation and decision-making.

Ethical Concerns Hard to Solve by Design alone

- Systemic & societal issues
  - Educational structures themselves may disadvantage disabled or marginalized students in ways design can’t fully fix.
- Unknown unknowns
  - Unforeseen problems will inevitably arise and need flexible, ongoing governance.
- Value change over time
  - Shifting ethical standards may invalidate original design assumptions.
- Transparency–privacy trade-off
  - Balancing explainability with data protection and fairness.

Something you want to add?

Moving on…

AI in Practice: What Can Go Wrong?

As AI systems move from lab environments to real-world applications, new kinds of ethical and practical problems emerge.

Why does this matter?

- AI systems may work well in testing but fail unpredictably in the real world.
- Ethical risks are often hidden in data, deployment context, or incentives.
- Failures can scale quickly and impact real lives
  - Discrimination, safety issues, loss of trust.
- Understanding how and why AI goes wrong is essential to preventing future harm.

What can go wrong in Practice?

- Bias and Discrimination
  - Biased training data leads to biased outputs. Replicates and reinforces social inequalities
  - Example: Hiring tools, predictive policing
- Lack of Transparency
  - Users don’t understand or cannot challenge AI decisions. Systems act as “black boxes”
  - Example: Credit scoring, medical diagnosis
- Overreliance / Automation Bias
  - People trust AI even when it is clearly wrong or misaligned with context
  - Example: GPS directions, autopilot
- Function Creep
  - AI used for one purpose expands silently into others (e.g., law enforcement use of commercial data)
  - Example: Smart speakers, surveillance

What can go wrong in Practice?

- Data Privacy Violations
  - Personal data collected without proper consent. Data being reused or sold
  - Example: Smart homes, mental health apps
- Unsafe Deployment
  - AI deployed in real-world contexts without sufficient testing or safeguards
  - Example: Self-driving cars, robotic surgery
- Feedback Loops
  - System learns from its own outputs, reinforcing narrow behavior (e.g., filter bubbles)
  - Example: Recommender systems, ad targeting

Why do these Failures happen?

- Biased Training Data
  - Data reflects past human biases (e.g., racist policing records, gendered job roles)
- Lack of Contextual Testing
  - Systems tested in narrow environments don’t generalize (e.g., from private roads to city streets)
- Misaligned Objectives
  - Optimizing for engagement, clicks, or efficiency can ignore fairness or well-being
- No Human Oversight
  - Systems make decisions without accountability mechanisms or intervention
- Incentive Misalignment
  - Companies optimize for speed or profit, not ethics or safety
- Lack of Regulation or Standards
  - Little to no legal limits on harmful deployment or poor design

Categories of AI Risk

- Epistemic risk
  - The system is wrong or misleading
  - Self-driving car misclassifies child as inanimate object

See “Hallucinations are mathematically inevitable”

“Large language models are increasingly relied upon as sources of information, but their propensity for generating false or misleading statements with high confidence poses risks for users and society.”

- Ghafouri et al. (2025) - Epistemic Integrity in Large Language Models (https://arxiv.org/abs/2411.06528)

Categories of AI Risk

Moral risk

- The system does harm or violates values
- Hiring AI favors privileged groups

1 *Discrimination & toxicity*	
1.1 Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
1.2 Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
1.3 Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.

Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence (https://arxiv.org/abs/2408.12622)

Categories of AI Risk

Political Risk (Power, Manipulation, Structural Bias)

- The system shifts power, often unequally
- Surveillance AI targets protesters. Data sold to authoritarian regimes

“Our findings reveal a complex landscape of potential threats, ranging from environmental harm and structural discrimination to governance failures and loss of control”

Uuk et al. (2024) - A Taxonomy of Systemic Risks from General-Purpose AI (https://arxiv.org/abs/2412.07780)

Categories of AI Risk

Social Risk (Behavioral / Relationship Change, Norms, Polarization)

- The system changes relationships or behaviors
- Recommender AI increases polarization and echo chambers

“Domain 5: Human-computer interaction
5.1 **Overreliance and unsafe use.** Users may come to trust or rely on AI systems beyond their actual capabilities or to anthropomorphize AI system [...]
5.2 **Loss of human agency and autonomy.** As AI systems become increasingly capable and intelligent, humans may be tempted to delegate many of their decisions and actions to AI (Paes et al., 2023)”

- Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence (https://arxiv.org/abs/2408.12622)

Categories of AI Risk

Legal Risk (Conflict With Law, Lack of Clarity, Compliance)

- Legal Risk (Conflict With Law, Lack of Clarity, Compliance)
- Chatbot offers health advice that violates medical regulations

"6.5 Governance failure. Governance failure refers to the risks and harms that arise when institutional, regulatory, and policy mechanisms fall short of effectively managing and overseeing the development and deployment of AI systems. Several issues make robust AI governance challenging to implement (Nah et al., 2023)."

- Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence (https://arxiv.org/abs/2408.12622)

Systemic Risk Categories

All of the following is taken from Uuk et al. (2024) - A Taxonomy of Systemic Risks from General-Purpose AI

Link: https://arxiv.org/abs/2412.07780

Systemic Risk Categories

- Control
  - The risk of AI models and systems acting against human interests due to misalignment, loss of control, or rogue AI scenarios.
- Democracy
  - The erosion of democratic processes and public trust in social/political institutions
- Discrimination
  - The creation, perpetuation or exacerbation of inequalities and biases at a large-scale.
- Economy
  - Economic disruptions ranging from large impacts on the labor market to broader economic changes that could lead to exacerbated wealth inequality, instability in the financial system, labor exploitation or other economic dimensions.

Systemic Risk Categories

- Environment
  - The impact of AI on the environment, including risks related to climate change and pollution.
- Fundamental Rights
  - The large-scale erosion or violation of fundamental human rights and freedoms.
- Governance
  - The complex and rapidly evolving nature of AI makes them inherently difficult to govern effectively, leading to systemic regulatory and oversight failures.
- Harms to non-humans
  - Large-scale harms to animals and the development of AI capable of suffering.

Systemic Risk Categories

- Information
  - Large-scale influence on communication and information systems, and epistemic processes more generally.
- Irreversible change
  - Profound negative long-term changes to social structures, cultural norms, and human relationships that may be difficult or impossible to reverse.
- Power
  - The concentration of military, economic, or political power of entities in possession or control of AI or AI-enabled technologies

Systemic Risk Categories

- Securit
  - The international and national security threats, including cyber warfare, arms races, and geopolitical instability.
- Warfare
  - The dangers of AI amplifying the effectiveness/failures of nuclear, chemical, biological, and radiological weapons.

Takeaways

- AI systems fail not only due to bugs, but also due to ethical blind spots.
- What works in lab conditions can be harmful in society.
- AI failures often reinforce existing social inequalities.
- Most failures are predictable and therefore preventable.
- Ethical foresight is part of responsible design, not idealism.