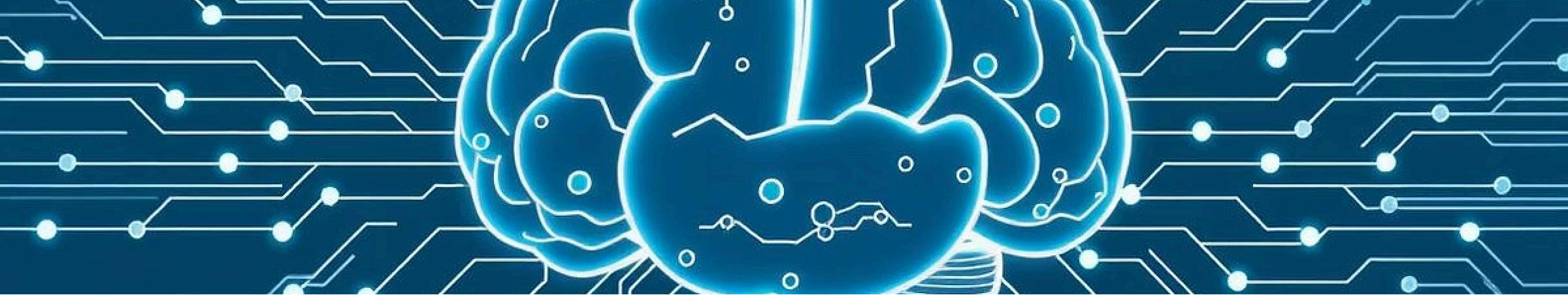


AI Ethics

Module 1: Introduction to AI Ethics

By Purva Thakare

A paradigmatic change

What Is a *Paradigmatic Change*?

Coined by philosopher Thomas Kuhn in *The Structure of Scientific Revolutions*, a **paradigmatic change** (or paradigm shift) occurs when the dominant way of understanding a field is replaced by a new framework that better explains or addresses challenges the old paradigm could not.

- From Ethical Principles to Power and Politics

Before: Focused on abstract principles like fairness and transparency.

Now: Ethics is increasingly viewed as a struggle over power, who benefits, and who is harmed.

Ethics becomes a political project, not just a technical or philosophical one.

Example: Questions about AI surveillance are no longer just about privacy, but about state power and civil liberties.

- From Universal Ethics to Contextual and Plural Ethics

Before: One-size-fits-all ethical models dominated, often rooted in Western norms.

Now: There is recognition that ethics must be plural, rooted in local, cultural, historical, and social contexts.

Example: What counts as “fair” AI in one culture might differ drastically in another — ethics must adapt to context.

- From AI Ethics as Tech Ethics to AI Ethics as Societal Ethics

Before: AI ethics focused on technical fixes for bias or safety.

Now: The shift is toward seeing AI as embedded in socioeconomic systems, impacting labor, democracy, inequality, and the environment.

Example: Debates about job automation are now ethical questions about economic justice.

- From Ethics by Design to Ethics by Governance

Before: Emphasis was on embedding ethics into systems at the design phase.

Now: Emphasis is also on institutional governance, public oversight, and legal accountability.

Example: The EU AI Act moves beyond ethical guidelines to binding law — ethics becomes enforceable.

- From Risk Mitigation to Structural Transformation

Before: AI ethics focused on avoiding harms like discrimination or misinformation.

Now: There's a deeper push to rethink the structure of the tech industry, including data extraction, corporate monopolies, and surveillance capitalism.

Example: Ethical AI is not just better facial recognition — it might mean banning facial recognition altogether.

AI Challenges

- **Bias and Fairness** AI models may reflect social, racial, or gender biases from training data.

Challenge: Ensuring fairness across diverse user groups.

- **Explainability and Transparency** Many AI systems (especially deep learning) are hard to interpret.

Challenge: Making AI decisions understandable and accountable.

- **Data Availability and Quality** AI needs large, high-quality datasets to learn effectively.

Challenge: Accessing, cleaning, and labeling data ethically and accurately.

- **Robustness and Safety** AI can fail in unpredictable or adversarial conditions.

Challenge: Building resilient, error-resistant systems.

- **Scalability and Integration** AI must scale across different industries and systems.

Challenge: Adapting AI to real-world, complex environments.

- **Privacy Protection** AI often depends on personal data for training and performance.

Challenge: Preserving user privacy while maintaining model performance.

- **Regulation and Governance** Global standards and legal frameworks are still evolving.

Challenge: Aligning innovation with ethical and legal requirements.

- **Energy and Environmental Impact** Training large models consumes massive energy.

Challenge: Developing sustainable and efficient AI systems.

Responsibility in the Ethics of Technology

Responsibility in the ethics of technology refers to the moral and practical obligations of individuals, organizations, and societies when developing, deploying, or using technological systems — especially those with wide-reaching impact like AI, biotechnology, or surveillance tools.

Types of Responsibility

- **Moral Responsibility** Who is ethically accountable when a technology causes harm?

Applies to designers, developers, users, and even policymakers.

Involves foresight, care, and a commitment to human values.

Example: Developers are morally responsible if an AI hiring tool unfairly discriminates against applicants.

- **Legal Responsibility** Concerns who is legally liable for harm caused by technology.

Legal systems often lag behind technological innovation.

Ongoing debates about liability for autonomous vehicles, AI decisions, etc.

- **Social Responsibility** Refers to the duty of tech companies and governments to serve the public good.

Includes transparency, inclusion, environmental impact, and preventing misuse.

Example: Ensuring access to technology for marginalized communities.

- **Professional Responsibility** Engineers, data scientists, and tech workers have codes of conduct.

Includes being honest about risks, protecting privacy, and reporting unethical practices.

Impact of AI/ML on Individuals & Society

Positive Impacts:

- Personalized healthcare diagnostics.
- Smarter education tools for individualized learning.
- Accessibility improvements for disabled communities.

Negative Impacts:

- Large-scale surveillance infringing on freedoms.
- Spread of misinformation and fake content.
- Widening socio-economic inequalities.

Societal Transformation:

- AI influencing employment patterns, cultural norms, and governance.

AI Ethical Frameworks & Implications

1

Key Ethical Principles: Fairness, accountability, transparency, privacy, safety.

2

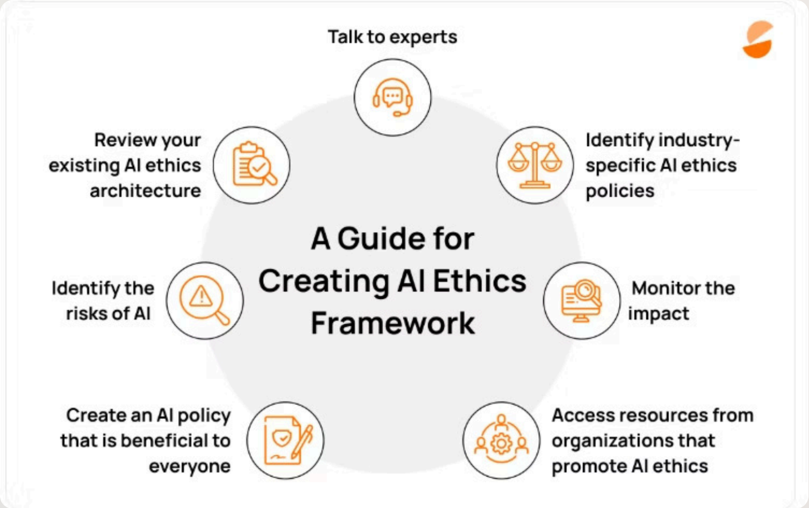
Implementation Methods:

- Bias testing and mitigation strategies.
- Explainable AI models for user trust.
- Regular ethical audits of AI systems.

3

Implications:

- Stronger public trust in AI applications.
- Prevention of harmful societal consequences.
- Long-term sustainable technological growth.



A Guide for Creating AI Ethics Framework

Talk to experts
Identify industry-specific AI ethics policies
Monitor the impact
Access resources from organizations that promote AI ethics
Create an AI policy that is beneficial to everyone
Identify the risks of AI
Review your existing AI ethics architecture

Cont

Core principles of AI ethical frameworks

Despite slight variations in different frameworks, there is a broad consensus around these core principles:

- **Fairness and non-discrimination:** AI systems should be developed to avoid and mitigate biases in training data and algorithms, preventing discriminatory outcomes based on characteristics like race, gender, or socioeconomic status.
- **Transparency and explainability:** The decision-making process of AI systems, particularly "black box" models, must be clear and understandable to users and regulators. This builds trust and allows for accountability.
- **Accountability and responsibility:** Clear lines of responsibility must be established for the outcomes of AI systems, since AI itself cannot be held liable. Human oversight and governance mechanisms are crucial.
- **Privacy and data protection:** With AI systems' reliance on vast amounts of data, frameworks emphasize safeguarding personal information and securing it from unauthorized access.
- **Human-centric values:** AI should augment human capabilities, not replace or dictate them, ensuring human oversight and control remain paramount in critical applications.
- **Beneficence and non-maleficence:** AI systems should be designed to maximize benefits to individuals and society while minimizing the risk of causing harm.
- **Safety and reliability:** AI systems must be robust, secure, and rigorously tested to prevent unintended failures or exploitation.

International frameworks:

- **UNESCO Recommendation on the Ethics of AI:** Adopted by 193 member states in 2021, this first-of-its-kind global agreement provides guidance on AI's ethical development and use.
- **OECD AI Principles:** The Organisation for Economic Co-operation and Development's principles focus on fostering innovation and public trust through inclusivity, human rights, and accountability.
- **EU AI Act:** The European Union's comprehensive AI law classifies AI systems by risk level and requires greater regulation for high-risk applications.

Corporate frameworks:

- **IBM:** Its principles emphasize that AI should augment human intelligence, and data and insights belong to the creator.
- **Microsoft and Google:** Both companies have internal ethics teams dedicated to promoting responsible AI development.

Non-profit organizations and research institutes:

- **AI Now Institute:** A research center at New York University that examines the social consequences of AI.
- **The Algorithmic Justice League (AJL):** This organization uses media and art to highlight the potential harms of algorithmic bias.

Challenges of implementing AI ethical frameworks

Despite growing consensus on core principles, implementing them effectively faces significant challenges.

- **Lack of universal standards:** There is no single globally accepted framework for AI ethics, creating a fragmented landscape of standards and priorities across different countries, organizations, and cultures.
- **Rapid technological advancement:** The pace of AI development often outstrips the ability of guidelines and regulations to keep up, leading to gaps in oversight.
- **Difficulty in defining ethical principles:** Abstract principles like "fairness" can be difficult to translate into concrete technical requirements, and there can be trade-offs between different ethical goals.
- **Bias in data and algorithms:** AI systems learn from existing data, which can perpetuate and even amplify historical societal biases if not carefully audited and addressed.
- **Lack of interdisciplinary collaboration:** Effective AI ethics requires input from diverse fields, including technology, law, and social sciences. However, cross-disciplinary collaboration is often limited.
- **Enforcement gaps:** Many ethical frameworks are voluntary and lack enforcement mechanisms or accountability measures, which can lead to superficial "ethics washing".
- **The "black box" problem:** The complexity of some AI models, particularly deep learning, can make it difficult to explain their decisions, hindering transparency and accountability.

Implications for society and technology

The ethical considerations and frameworks for AI have profound implications for both society and the technology itself.

Technological implications:

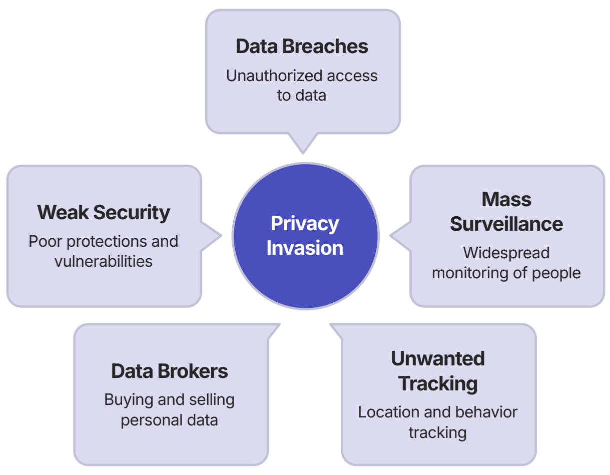
- **Shift toward responsible design:** Ethical frameworks encourage developers to integrate ethical principles from the initial design phase ("Ethics by Design"), rather than as an afterthought.
- **New technological tools:** The need to address ethical challenges is spurring innovation in technologies that can, for example, detect and mitigate bias, increase transparency, and protect privacy.
- **Standardization and governance:** Ethical frameworks are driving the development of international and industry-specific standards for AI governance, leading to greater consistency and predictability.

Societal implications:

- **Bias and inequality:** Unethical AI can exacerbate existing inequalities through biased hiring algorithms, discriminatory credit scoring, or unfair judicial systems.
- **Accountability and liability:** Assigning responsibility for AI-induced harm is a complex legal and ethical challenge, particularly with autonomous systems like self-driving cars.
- **Loss of privacy and autonomy:** Pervasive data collection and AI-driven surveillance can infringe on personal privacy and potentially manipulate human behavior.
- **Existential risks:** In the long term, the unchecked development of highly intelligent AI could pose significant risks to humanity, requiring robust ethical guardrails.
- **Public trust and adoption:** Adherence to ethical frameworks is crucial for building and maintaining public trust, which is necessary for the successful adoption of AI technologies.
- **Workforce impact:** AI's potential to displace jobs and create new ones raises ethical questions about labor, income inequality, and the need for new economic models.

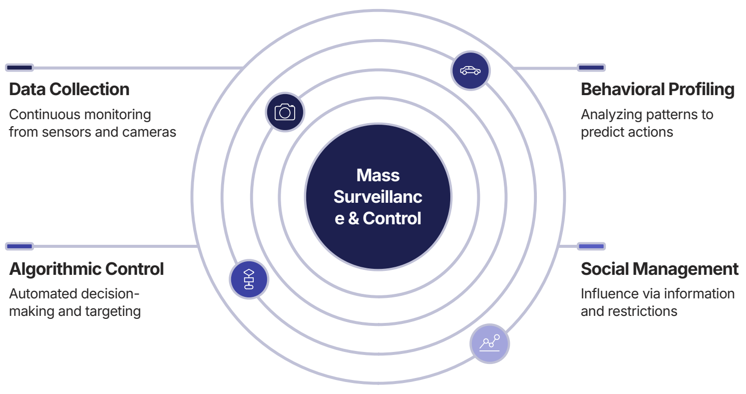
Surveillance Issues in AI

Privacy Invasion


Data Breaches
Unauthorized access to data
Weak Security
Poor protections and vulnerabilities
Privacy Invasion
Mass Surveillance
Widespread monitoring of people
Data Brokers
Buying and selling personal data
Unwanted Tracking
Location and behavior tracking

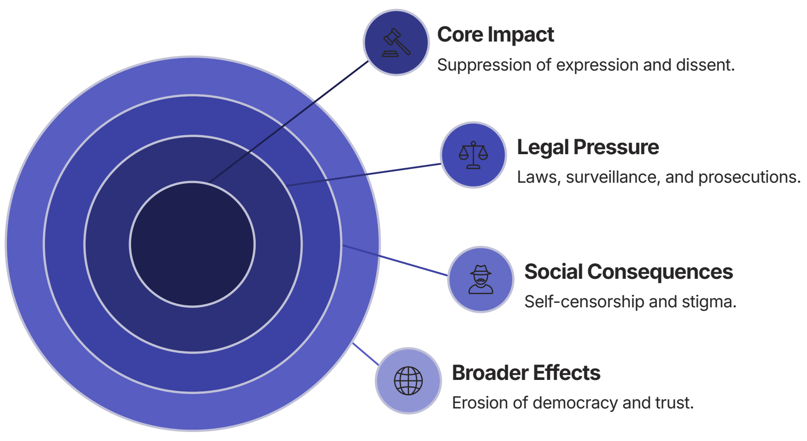
- **Description:** AI systems like facial recognition and online trackers collect personal data without explicit consent, reducing people's control over their information.
- **Example:** Smart CCTV cameras in Beijing and London monitor citizens' movements, even when they haven't committed any crime.

Mass Surveillance & Control


Data Collection
Continuous monitoring from sensors and cameras
Behavioral Profiling
Analyzing patterns to predict actions
Mass Surveillanc e & Control
Algorithmic Control
Automated decision-making and targeting
Social Management
Influence via information and restrictions

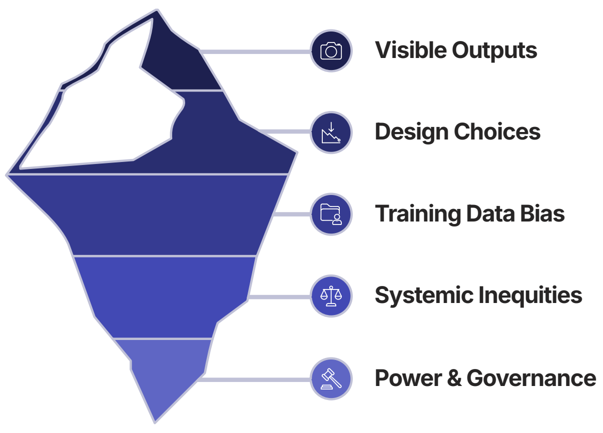
- **Description:** Governments or corporations use AI to monitor populations at scale, potentially enabling authoritarian control and social manipulation.
- **Example:** China's **social credit system** tracks behaviors (like purchases, debts, political opinions) to reward or punish citizens.

Chilling Effect on Freedom


Core Impact
Suppression of expression and dissent.
Legal Pressure
Laws, surveillance, and prosecutions.
Social Consequences
Self-censorship and stigma.
Broader Effects
Erosion of democracy and trust.

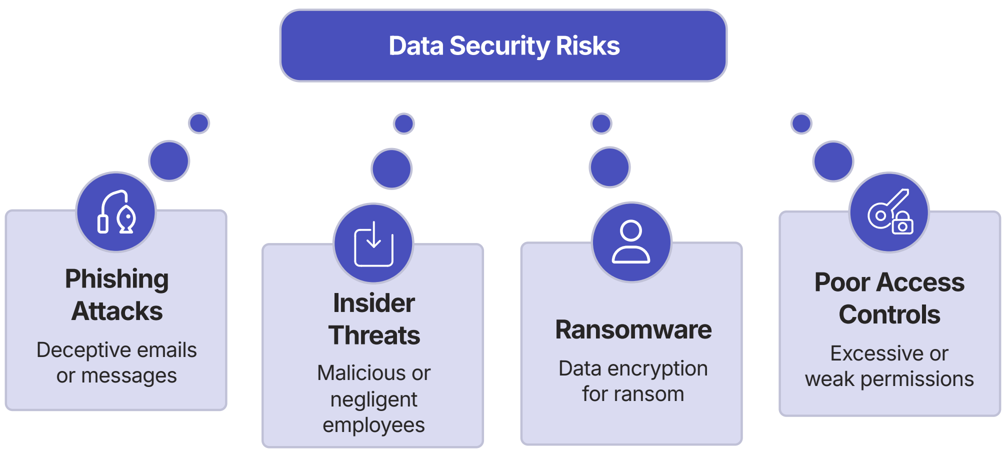
- **Description:** Knowing one is constantly watched discourages free speech, activism, and protests.
- **Example:** Protesters in Hong Kong wore masks to avoid being identified by AI-powered surveillance cameras.

Bias in Surveillance Tools



- **Description:** AI surveillance often shows racial or gender bias, leading to false identifications and unfair treatment.
- **Example:** **Amazon's facial recognition tool** was found to misidentify Black women more frequently, raising risks of wrongful arrests.

Data Security Risks


Data Security Risks
Phishing Attacks
Deceptive emails or messages
Insider Threats
Malicious or negligent employees
Ransomware
Data encryption for ransom
Poor Access Controls
Excessive or weak permissions

- **Description:** Data collected through AI surveillance can be hacked, leaked, or misused, violating individual rights.
- **Example:** If health data from AI-powered fitness trackers is stolen, employers or insurers may exploit it to discriminate.

Segmentation Issues in AI

Discrimination & Bias

Description: AI systems may segment people unfairly based on gender, race, or background, reinforcing inequalities.

Example: Amazon’s hiring AI discriminated against women because it learned patterns from male-dominated historical data.

Digital Redlining

Description: Certain groups may be excluded from opportunities because AI segments them into less “profitable” or “desirable” categories.

Example: Facebook’s ad-targeting algorithm once prevented minorities from seeing housing ads, violating fair housing laws.

Overgeneralization

Description: AI places people into broad categories, ignoring individual differences.

Example: Car insurance AI may charge higher premiums to all residents of a “high-crime” area, even if an individual has a perfect driving record.

Lack of Transparency

Description: People often don’t know why AI assigned them to a specific category or segment.

Example: YouTube recommends certain political videos after segmenting users (e.g., “likely conservative”) without disclosing the reasoning.

Manipulation of Autonomy

Description: AI segmentation can be used to exploit personal weaknesses and manipulate behavior.

Example: Cambridge Analytica used Facebook data to segment voters and send targeted political ads, influencing elections.

Challenges of AI Surveillance

ⓘ 1 **Balancing Security vs. Privacy**

Challenge: Governments argue surveillance is needed for safety, but it often comes at the cost of individual privacy.

Example: Airport facial recognition systems help security but scan millions of innocent passengers too.

ⓘ 2 **Regulation & Legal Boundaries**

Challenge: Lack of clear global laws on what level of AI surveillance is acceptable.

Example: The EU's GDPR restricts mass tracking, but many countries have no such rules.

ⓘ 3 **Bias and Wrongful Identifications**

Challenge: Surveillance tools misidentify minorities, increasing chances of wrongful arrests or discrimination.

Example: US police departments have faced criticism for relying on biased facial recognition matches.

ⓘ 4 **Overreach by Authorities**

Challenge: Surveillance data can be used for political suppression.

Example: Activists in some countries are tracked and silenced using AI monitoring tools.

ⓘ 5 **Data Storage & Security Risks**

Challenge: Sensitive data collected must be stored securely—but breaches are common.

Example: Data leaks from health tracking apps expose private information to hackers.

Challenges of AI Segmentation

1

Fairness and Discrimination

- Challenge: Algorithms often reinforce existing inequalities by categorizing based on biased historical data.
- Example: Women being underrepresented in job recommendations due to biased training data.

2

Transparency and Explainability

- Challenge: People don’t know why an AI assigned them to a segment → "black box" decision-making.
- Example: A loan applicant being denied credit without explanation of why they were “high risk.”

3

Exploitation of Vulnerable Groups

- Challenge: AI segmentation can be misused to manipulate people’s behavior.
- Example: Political ads targeting emotionally vulnerable users during elections.

4

Exclusion and Inequality

- Challenge: AI can digitally “redline” communities by excluding them from housing, loans, or healthcare opportunities.
- Example: AI in healthcare allocating fewer resources to low-income areas.

5

Accountability & Responsibility

- Challenge : When an AI segmentation system causes harm, it’s often unclear **who should be blamed**—the company deploying it, the developers who built it, or the data providers who trained it.

Example:

- In 2019, **Apple’s credit card (Apple Card)** faced criticism because women reportedly received **much lower credit limits** than men, even with better financial histories.
- The segmentation algorithm (developed by Apple and Goldman Sachs) appeared biased.
- But when complaints arose, **Apple blamed Goldman Sachs (the issuer)**, Goldman Sachs blamed the **AI system’s design**, and developers blamed the **training data**.
- This created confusion: **Who was actually responsible for the harm?**