Module III:

AI Code of Ethics (Privacy,Governance Policy, and Transparency)

AI Policy

What are AI Policies?

AI policies serve as a guiding framework for organizations, delineating the principles, guidelines, and procedures **governing** the deployment and use of **AI systems**. These policies are crafted to align with legal requirements, ethical standards, and organizational values, ensuring that AI technologies are used responsibly and **ethically**.

Key Components of an AI Policy

1. Introduction: An AI policy typically begins with an introductory section outlining the organization's commitment to compliance with relevant laws and ethical use of AI. This section also highlights the importance of AI in driving innovation and enhancing organizational capabilities.

2. Purpose: The purpose section clarifies the objectives of the AI policy, emphasizing the need to establish clear guidelines for the responsible and ethical use of AI technologies within the organization. It aims to ensure alignment with legal requirements, ethical standards, and organizational values.

3. Scope: This section defines the scope of the AI policy, specifying the applicability of the policy to all AI-related activities within the organization. It outlines the boundaries within which the policy operates and clarifies the types of AI technologies and applications covered under the policy.

4. Definitions: Clear definitions of AI-related terms are crucial for ensuring common understanding across the organization. This section clarifies terminology such as "artificial intelligence," "AI system," "embedded AI tools," and other relevant terms to avoid ambiguity and confusion.

5. Guiding Principles: The heart of an AI policy lies in its guiding principles, which articulate the organization's stance on AI usage. These principles underscore the importance of ethical considerations, legal compliance, transparency, and accountability in AI deployment. Additionally, they may emphasize the organization's commitment to diversity, equity, and inclusion in AI development and deployment.

6. Prohibited Uses: This section delineates activities strictly prohibited in AI usage, such as conducting political lobbying, categorizing individuals based on protected class status, or entering sensitive information into AI systems. Organizations mitigate risks by explicitly outlining prohibited uses and ensuring alignment with legal and ethical standards.

7. Ethical Guidelines: While some AI applications may be legally permissible, they might not align with ethical standards. Ethical guidelines ensure that AI usage upholds principles of informed consent, integrity, appropriateness, and respect for privacy. Organizations may also incorporate fairness, accountability, and transparency principles into their ethical guidelines to promote responsible AI development and deployment.

8. High-Risk Use of AI Systems: Certain AI applications pose heightened risks to individuals' rights and safety. This section outlines additional requirements and safeguards for high-risk AI applications, such as personnel decisions, job screening, or student assessments. Organizations must adhere to stringent criteria to mitigate risks and ensure compliance with regulatory requirements.

9. Reporting Non-Compliance: A robust reporting mechanism encourages employees to report violations or concerns regarding AI usage without fear of reprisal. This section outlines the reporting process, including channels for reporting, confidentiality measures, and protections against retaliation.

Principles and methods of ensuring differential privacy and data, regulations, and Trust-worthy AI.

What is data privacy?

Before discussing the intricacies of differential privacy, let's first establish a fundamental understanding of data privacy. **Data privacy** encompasses safeguarding personal and sensitive data from unauthorized access, disclosure, or misuse, empowering individuals to regulate who can access their personal information. Various measures, such as the General Data Protection Regulation (GDPR), the CCPA (California Consumer Protection Act), and the Health Insurance Portability and Accountability Act (HIPAA), are implemented to uphold individuals' privacy rights and impose rigorous regulations on organizations handling personal data. Consequently, organizations are bound by legal obligations to maintain high data protection and anonymization standards.

The exact definition of personal data varies depending on specific laws in different countries or regions, but it typically covers any information that relates to an individual, including personally identifiable information (**PII**), obvious confidential information, biometric data, geolocation data, internet usage data, and online identifiers.

Data privacy is essential for several reasons, among them:

- **Upholding fundamental rights:** it is a fundamental right that protects personal data and upholds freedom in an increasingly interconnected digital world.
- **Protecting personal information:** it is essential to maintain the confidentiality of sensitive information, such as names, addresses, financial details, and health records.
- **Preventing data breaches and cyberattacks:** its measures can protect personal information from being leaked and prevent malicious hackers and cybercriminals from targeting personal data for fraud, identity theft, and other crimes.

Differential Privacy

Differential privacy (DP) is a mathematical framework that ensures individuals cannot be identified within a dataset, even when releasing aggregate information. It is not a specific technique but a quantifiable property that an algorithm or process can possess. DP prevents re-identification by adding controlled, random "noise" to the data, ensuring that statistical outcomes remain approximately the same whether a single person's data is included or not.

Key Principles:

Indistinguishability: An algorithm is differentially private if its output is almost identical whether or not an individual's data is included. An outside observer cannot confidently determine if any single individual was part of the original dataset.

Privacy budget (): DP measures the trade-off between data utility and privacy using a parameter called epsilon (). A smaller $\epsilon epsilon$ () offers stronger privacy protection but adds more noise, reducing accuracy. A larger () provides less privacy but preserves higher data accuracy.

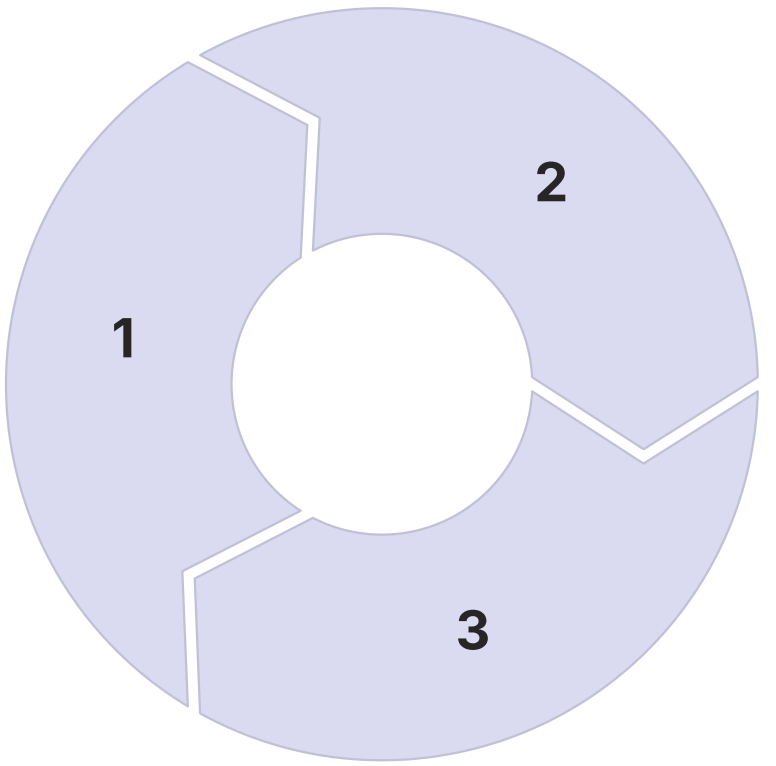
Composition: The total privacy loss accumulates over multiple analyses of the same dataset. The principle of composition allows you to track and manage this cumulative privacy loss (or privacy budget) across multiple queries.

Robustness to post-processing: Any computation performed on the output of a differentially private algorithm cannot reduce its privacy guarantees. An attacker cannot use outside information to reverse-engineer and compromise the privacy of the original data.

Methods for ensuring differential privacy

Noise injection mechanisms

Laplace Mechanism: This is used for numerical data and adds noise drawn from a Laplace distribution to the result of a query. The amount of noise is proportional to the query's sensitivity and the privacy budget.

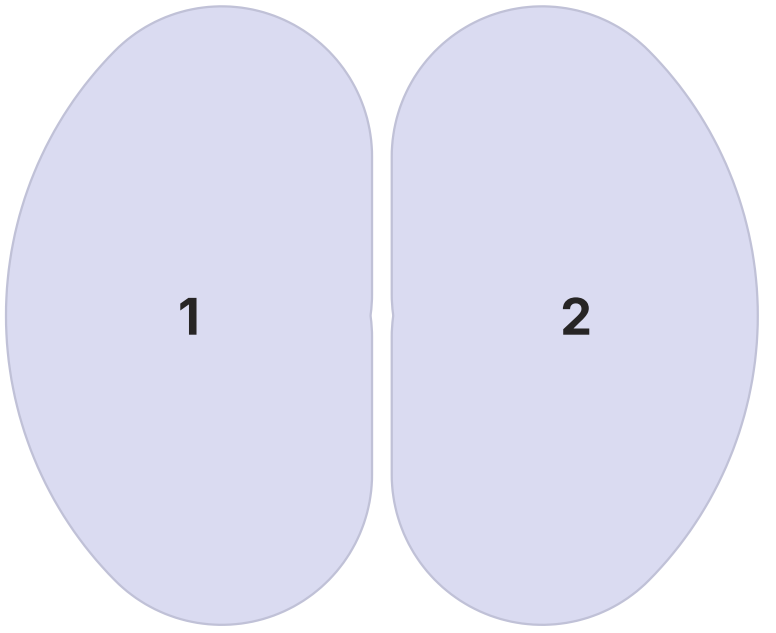


Gaussian Mechanism: Similar to the Laplace mechanism but uses a Gaussian (normal) distribution to add noise. It is often used for algorithms that require a different measure of sensitivity.

Randomized Response: For sensitive, non-numerical data (like "yes/no" survey questions), individuals can randomly decide to tell the truth or lie based on a coin flip. This adds noise at the individual level (Local DP), ensuring plausible deniability.

Advance Methods

Federated Learning with DP: Training a machine learning model across decentralized devices (like mobile phones) without centralizing the raw user data. Differential privacy is applied to the model updates sent from each device to prevent a central server from learning about any specific user.



Differential Privacy Stochastic Gradient Descent (DP-SGD): A method used to train deep learning models with differential privacy by adding noise to the gradients during the training process.

Regulatory examples and real-world applications

US Census Bureau (Government)

- **Regulation:** The US Census Bureau adopted differential privacy for the 2020 Decennial Census to protect respondents' confidentiality while releasing statistical data for redistricting.
- **Application:** To prevent the re-identification of individuals, the Census Bureau added noise to the aggregated data. For example, a census table showing the exact number of people with a specific characteristic in a small town would be altered slightly. The released statistic might show "24 people" instead of the true "23 people," making it impossible to pinpoint any single person's information while still providing a highly accurate aggregate picture.

Apple (Technology)

- **Regulation:** Apple uses differential privacy to collect user data while complying with global privacy laws.
- **Application:** Apple uses Local Differential Privacy to gather insights about user behavior, such as which emojis are most popular, which websites cause performance issues in Safari, and which words are added to a user's local dictionary. The noise is added on the user's device before data is sent to Apple, ensuring that Apple never receives an individual's true data point.

Google (Technology)

- **Regulation:** Google complies with privacy regulations by employing differential privacy in many products.
- **Application:** Google uses DP to collect browser statistics in Chrome and to power other services. Its open-source differential privacy libraries are available to developers to help them build their own privacy-preserving applications.

Financial Institutions (Banking)

- **Regulation:** Banks must comply with regulations protecting sensitive customer financial data.
- **Application:** To analyze transaction patterns without exposing individual customers, a bank could use differential privacy. A bank could calculate the average daily transaction amount in a region by adding noise to the aggregated result. This protects the data of any individual high-spending customer from being isolated, while still allowing the bank to understand general trends for fraud detection.

Accountability and transparency of AI

Accountability and transparency in AI involve **establishing responsibility for AI outcomes and making AI systems understandable to humans**. Accountability ensures that people or organizations are answerable for AI's actions, while transparency allows for an understanding of how the system works, its data, and its decision-making processes. Key strategies include using explainable AI (XAI), establishing clear governance and regulatory frameworks, conducting regular audits, and promoting a culture of responsible development and deployment.

Accountability

- **Define roles and responsibilities:** Clearly state who is responsible for the design, development, and deployment of AI systems.
- **Establish ownership:** Create frameworks to ensure that individuals and organizations are answerable for the outcomes and impacts of the AI.
- **Implement governance:** Develop governance structures that define roles and duties, manage risks, and oversee the AI lifecycle.
- **Conduct validation and audits:** Regularly audit AI systems to evaluate their effectiveness, fairness, and ethical implications, and correct errors or biases.

Transparency

- **Explain AI systems:** Make the internal workings of AI systems, including their algorithms and decision-making processes, understandable to users and stakeholders.
- **Communicate data usage:** Be clear about the data used to train the AI and how it is collected and processed.
- **Use explainable AI (XAI):** Implement XAI methods to provide insight into why an AI system made a particular decision.
- **Foster open communication:** Create open communication channels among developers, users, and regulators about the system's capabilities and limitations.

Why they are crucial

- **Build trust:** Transparency and accountability are essential for building public and stakeholder trust in AI technologies.
- **Prevent harm:** They help prevent and mitigate potential harm, such as biases in loan applications, misdiagnoses in healthcare, or other negative consequences.
- **Meet regulations:** Transparency is a growing legal requirement in many areas, such as with the EU's AI Act.

Responsibility and Accountability related to AI Design and Deployment

Responsibility in AI Definition:

Responsibility refers to the ethical and professional duty of individuals or organizations involved in creating and using AI systems to ensure they are safe, fair, transparent, and beneficial to society.

Key Aspects of Responsibility:

Ethical Design and Development:

AI designers and developers must ensure the system follows ethical principles such as fairness, non-discrimination, transparency, and human well-being.

Data Responsibility

Use high-quality, unbiased, and lawful data.

Respect privacy and consent during data collection and training.

Algorithmic Transparency:

AI systems should be explainable — users should understand how decisions are made.

Human Oversight

AI should support, not replace, human judgment.

Humans must retain the ability to override or audit AI decisions when necessary.

Continuous Monitoring

Even after deployment, AI models should be regularly tested and updated to prevent unintended consequences.

Accountability in AI

Definition:

Accountability means that clear responsibility must be assigned for the outcomes of AI systems — whether positive or negative. If an AI system causes harm or error, it should be possible to trace who is answerable: developer, deployer, or organization.

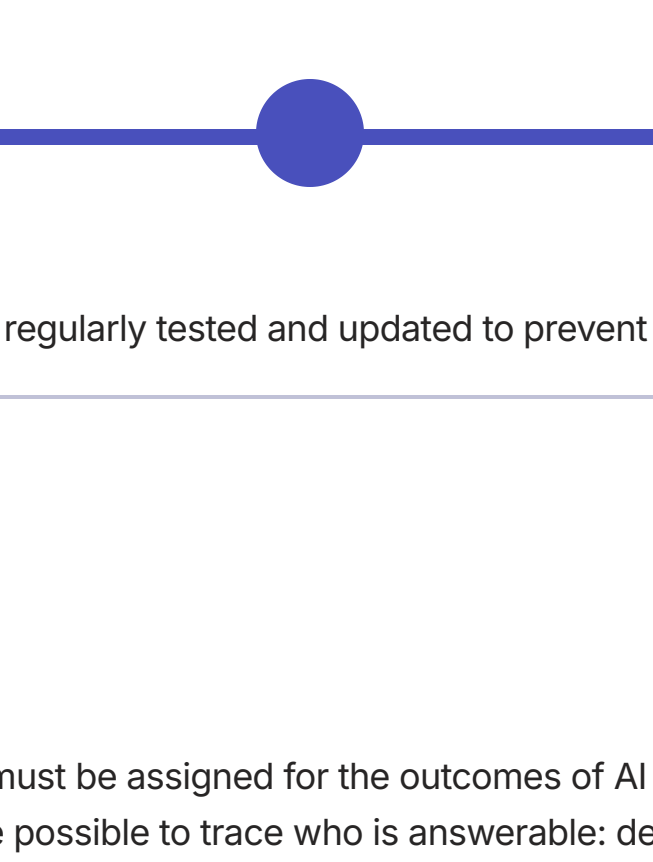
Key Aspects of Accountability:

Traceability:

Every stage of AI development (data collection, training, deployment) must be **documented** so that decisions can be reviewed and explained.

Human-in-the-loop:

Keeps humans accountable for final decisions in critical areas (e.g., hiring, medical diagnosis, sentencing).



Liability: Clear **legal responsibility** for damages caused by AI — e.g., if an autonomous car causes an accident, laws must specify whether the manufacturer or software provider is liable.

Auditing and Reporting:

Regular **independent audits** ensure compliance with ethical and legal standards.

Governance and Regulation: Organizations should adopt **AI governance frameworks** (e.g., EU AI Act, OECD principles, NITI Aayog's Responsible AI in India).

AI transparency and the changing nature of work and organizing

AI transparency: The foundation for trust and ethics

AI transparency involves providing clear explanations of how AI models function, the data they use, and how their decisions are made. This is essential for addressing the "black box" problem of complex algorithms that are inherently difficult to interpret.

The key pillars of AI transparency are:

- **Explainability:** The ability to describe to non-experts how an AI system reached a decision.
- **Interpretability:** Focuses on the inner workings of a model, showing how specific inputs lead to outputs.
- **Accountability:** Ensures that AI systems and the organizations deploying them are held responsible for outcomes and errors.

The importance of transparency increases with the stakes of the AI application. For low-impact systems like product recommendations, less transparency may be acceptable. However, for "high-stakes" applications in finance, healthcare, or human resources, greater transparency is critical to ensure fairness and prevent significant negative impacts.

The changing nature of work

AI is creating a shift in the workplace that goes beyond simple automation. The changes are redefining job roles, shifting skill requirements, and requiring new forms of human-machine collaboration.

Impact on jobs:

- **Task automation:** AI is automating repetitive and routine tasks, freeing human workers to focus on more complex, high-value activities that require creativity, critical thinking, and social and emotional intelligence.
- **Job augmentation:** In many roles, AI does not replace workers but instead acts as a "superpower" that augments human capabilities, making them more productive and effective.
- **New job categories:** While some jobs are at risk of displacement, AI is creating entirely new job categories, such as AI workflow optimizers and roles focused on managing and training AI systems.
- **Shifting skills:** The emphasis is moving from traditional, task-based skills toward human-centric ones like communication, problem-solving, and adaptability.

Challenges for workers:

- **Anxiety and resistance:** Many employees fear job loss or disruption from AI implementation, which can lead to resistance if not addressed proactively.
- **Discomfort with AI:** Studies show that increased transparency can increase employee discomfort with AI because it makes the systems seem more human-like, challenging human uniqueness.
- **Bias and fairness:** Without transparency and oversight, AI systems can amplify existing biases in their training data, leading to discriminatory outcomes in hiring or performance management.
- **Data privacy and surveillance:** The data collection required for AI raises significant concerns about employee privacy and potential surveillance.

The transformation of organizing and management

Integrating AI successfully requires a fundamental shift in organizational culture and management practices.

New organizational imperatives:

- **Strategic change management:** Companies must adopt a people-centric approach to change management, involving employees early in the AI journey to reduce resistance and foster a culture of adaptability.
- **Emphasis on human-AI teams:** The new organizational model is one of human-machine partnership, where workflows are redesigned to leverage the complementary strengths of humans and AI.
- **AI ethics and governance:** Organizations must develop clear AI governance frameworks and ethics policies to ensure AI is developed and deployed responsibly. This includes establishing accountability mechanisms and conducting regular audits.
- **Redefining leadership:** The power dynamic in organizations is shifting from traditional operational management toward those with the foresight to leverage AI for data-driven strategy.

The role of transparent organizing:

- **Building trust with stakeholders:** External transparency, such as clear communication about AI practices in sustainability reports, can build trust with customers, investors, and regulators.
- **Empowering employees:** Internal transparency about how AI is used can empower employees, increase their confidence, and encourage active participation in the transformation process.
- **Regulatory compliance:** Transparent practices are becoming a prerequisite for regulatory compliance, as seen with frameworks like the EU AI Act.

Human-in-the-loop AI

What is Human-in-the-Loop (HITL) decision-making?

Human-in-the-loop decision-making involves a human decision-maker working in tandem with AI algorithms to improve decision-making outcomes. In HITL systems, the AI system provides recommendations or predictions that the human decision-maker evaluates and approves or rejects.

HITL decision-making is used in a wide range of applications – as discussed later below, from health care diagnostics to decision-making software.

How does human-in-the-loop work?

Human-in-the-loop systems typically involve an algorithm that is trained to recognize patterns in data and produce an outcome, as well as a human decision-maker who works alongside the AI system (as opposed to a fully automated system).

First, the AI is given an input, such as data or a text prompt, and makes a preliminary decision. The human decision-maker can then correct or adjust the AI's recommendations based on their expertise, knowledge and understanding of the context.

This blended decision-making approach ensures greater efficiency in the decision-making process than fully human decision-making would allow, while achieving higher accuracy and more desirable outcomes than a fully automated decision-making process.

Designing Human-in-the-Loop systems

The design of HITL systems is a cyclical process that involves humans at various stages of an AI model's lifecycle.

1. **Data annotation and labeling:** Human experts label and categorize initial training data, providing the foundational "ground truth" for the AI model to learn from.
2. **Model training and evaluation:** During training and testing, humans review the AI's predictions and outputs, providing feedback that helps refine the model. This can involve:
    - **Active learning,** where the AI asks humans for input on the data points it is most uncertain about.
    - **Reinforcement Learning from Human Feedback (RLHF),** where humans rank or correct an AI's responses to optimize its performance.
3. **Real-time intervention:** In critical or complex situations, human operators can correct or override AI outputs before they are delivered. This is crucial for applications where the cost of error is high.
4. **Continuous feedback loop:** Even after deployment, human monitoring and feedback help the AI adapt to evolving needs and correct errors over time, ensuring it remains effective in real-world scenarios.

Benefits of Human-in-the-Loop AI

- **Improved accuracy and reliability:** Humans correct AI errors, especially in ambiguous or "edge cases" where an AI's programming falls short. This collaboration strengthens the model and builds user trust.
- **Ethical control and bias mitigation:** Human oversight is essential for identifying and rectifying biases inadvertently embedded in the training data, promoting fairness and preventing discriminatory outcomes.
- **Enhanced transparency and explainability:** By embedding human review, AI decisions become more transparent and easier to interpret. When an output is reviewed, a human can often provide the reasoning, demystifying the AI's "black box" nature.
- **Accountability:** Assigning final decision-making power to a human ensures that responsibility for the system's actions does not rest solely on the algorithm or its developers. This human accountability is legally and ethically crucial in sensitive areas like medicine or finance.
- **Adaptability:** Human contextual understanding allows the AI to remain flexible and adaptive to new situations, societal norms, and cultural nuances that algorithms alone would miss.

Challenges and considerations

Despite its benefits, implementing HITL systems presents several challenges:

- **Scalability:** Relying on human labor for constant review can be expensive, time-consuming, and create bottlenecks, especially as the volume of tasks increases.
- **Human error and inconsistency:** Human judgment is not infallible and can introduce its own biases and inconsistencies. If human reviewers are tired, distracted, or poorly trained, they can negatively impact the AI's learning process.
- **Automation bias:** People may over-rely on or uncritically accept an AI's output, diminishing the effectiveness of human oversight.
- **Defining the right level of control:** Striking the correct balance between machine autonomy and human intervention is critical and depends on the specific use case, risk level, and required speed.

Best practices for ethical design

To ensure ethical control and accountability, designers can follow these best practices:

- **Design for transparency:** Build user interfaces that make it clear what the AI is doing and why. Display the AI's confidence levels and allow users to see its reasoning to build trust and prevent automation bias.
- **Establish clear guidelines and governance:** Define ethical frameworks, establish clear roles, and create robust governance structures that detail when and how humans should interact with the system.
- **Train humans to work with AI:** Ensure that human operators have the necessary training to understand the AI's capabilities and limitations, so they can effectively monitor and intervene when needed.
- **Prioritize user control:** Give users easy-to-access mechanisms to override or reverse an AI's output. For high-stakes applications, this control is non-negotiable.
- **Implement continuous monitoring:** Establish ongoing feedback loops and audits to track the system's performance, identify emerging biases, and retrain the model with human-validated data.
- **Engage diverse stakeholders:** Involve a diverse group of people, including domain experts, users, and ethicists, throughout the design process to surface potential blind spots and ensure inclusivity.