Module IV:

Ethics Washing, Explainability, Robustness

Understanding Ethics Washing

What is Ethics Washing?

Ethicswashing refers to the practice where organizations or institutions **publicly emphasize ethical values or AI ethics initiatives** to appear responsible and trustworthy, while **failing to take meaningful action** to address real ethical problems in their technologies or operations. It is similar to *greenwashing* in environmental contexts, where companies overstate their environmental responsibility.

> **Ethics Washing = Saying "We're ethical" without *being* ethical.**

Why It Happens

Organizations engage in ethics washing to:

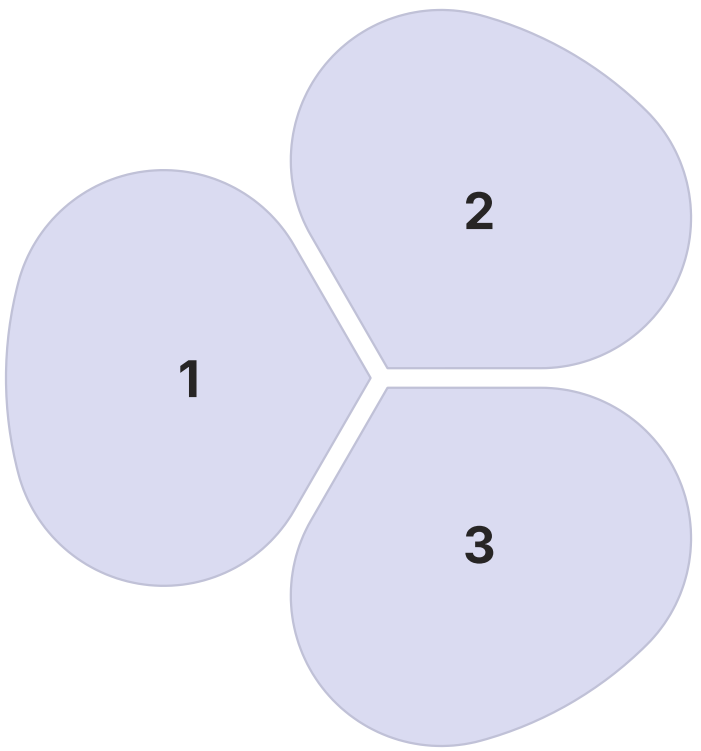
1. **Avoid regulation or criticism** while maintaining public trust.
2. **Promote their brand image** as "ethical" or "responsible."
3. **Delay accountability** by forming committees or issuing vague ethics principles without enforcement.

Examples

1. **AI companies creating ethics boards** that have no decision-making power or are dissolved when facing controversy.
2. **Tech firms publishing "AI principles"** but continuing practices like data exploitation or algorithmic bias.
3. **Corporations claiming "fair AI"** while keeping models and datasets secret, preventing external audits.

Consequences

Erosion of public trust in ethical AI efforts.



Slower progress toward genuine responsible AI development.

Misleading policymakers and users about real ethical risks.

How to Avoid Ethicswashing

1. **Transparency:** Share data sources, model details, and evaluation methods.
2. **Accountability:** Ensure ethical principles are enforced through audits and independent oversight.
3. **Stakeholder inclusion:** Involve affected communities, ethicists, and diverse experts in decisions.
4. **Action over words:** Demonstrate concrete ethical practices — not just policies.

Explainability from a user perspective: Explainability problem

What is explainable AI?

Explainable **artificial intelligence** (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by **machine learning** algorithms.

Explainable AI is used to describe an AI model, its expected impact and potential biases. It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making. Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production. AI explainability also helps an organization adopt a responsible approach to AI development.

What Is Explainability?

Explainability in **artificial intelligence** refers to the ability to describe an AI model's internal workings or outcomes in understandable terms. It makes complex AI decisions transparent and trustworthy. In fields like healthcare or finance, where understanding why a model made a particular decision has implications, explainability has influence. In terms of MLOps and **AI security**, explainability supports accountability and helps diagnose and rectify model errors

Why Explainability Matters

Machine learning models, particularly those based on complex algorithms like neural networks, can act as black boxes, obscuring the if/then logic behind their outputs. This opacity can lead to mistrust or skepticism among stakeholders, regulators, and customers who need to understand the basis of decisions impacting them.

In healthcare, for example, an AI system could be employed to assist radiologists by prioritizing cases based on the urgency detected in X-ray images. In addition to performing with high accuracy, the AI system must provide explanations for its rankings to ensure patient safety and comply with medical regulations. In other words, it needs to be transparent enough to reveal the features in the images that led to its conclusions, enabling medical professionals to validate the findings.

Additionally, in jurisdictions with regulations such as the EU's **General Data Protection Regulation (GDPR)**, patients may have the right to understand factors influencing their cases and could challenge decisions made with the aid of AI. In instances such as this, explainability goes beyond technical performance to encompass legal and ethical considerations.

Basic Problems in Making Algorithms Explainable

- **Accuracy vs. Interpretability Trade-off:** More complex, highly accurate models (like deep neural networks) are generally less interpretable, while simpler, transparent models (like linear regression or decision trees) often sacrifice predictive performance on complex data.
- **Model Complexity:** Modern machine learning systems can have millions of parameters and intricate internal workings, making it difficult to trace exactly how a specific input leads to a particular output, even for the developers themselves.
- **Lack of Standardized Evaluation Metrics:** There is no universal consensus or objective metric for what constitutes a "good" explanation. Different stakeholders (e.g., a data scientist, a regulator, an end-user) require different types and levels of explanation, making standardized evaluation difficult.
- **Human Bias:** Explainability methods can be affected by biases present in the training data or the design choices made by human developers. Explanations might inadvertently reinforce or obscure these biases, leading to unfair or discriminatory outcomes.
- **User Understanding:** Explanations must be tailored to the target audience's expertise. A technical explanation suitable for an ML engineer may be incomprehensible to a domain expert (e.g., a doctor or a loan officer), which can lead to over-reliance (automation bias) or under-reliance (algorithmic aversion) on the AI system.
- **Computational Expense:** Many post-hoc explainability methods, such as SHAP or LIME, are computationally expensive, especially for large models or real-time applications, which can limit their practicality.
- **Causality vs. Correlation:** Most current XAI methods highlight correlations and feature importance, but they struggle to provide true causal explanations, which is often what humans need to make informed decisions and act effectively.

Approaches to Making Algorithms Explainable

1

Intrinsically Interpretable Models (Explainable by Design): These models are designed from the ground up to be transparent and their operations are directly understandable by humans.

- **Linear/Logistic Regression:** The coefficients directly indicate the weight of each feature's influence on the output.
- **Decision Trees/Rule-based Systems:** The logic follows explicit "if-else" conditions that are easy to follow and understand.
- **Generalized Additive Models (GAMs):** These models allow the impact of each feature to be visualized individually.

2

Post-Hoc Explanation Techniques (for "Black-Box" Models): These methods are applied after a complex model (like a neural network or ensemble method) has been trained to provide insights into its behavior.

- **Techniques:**
  - **Feature Relevance Explanations (e.g., SHAP, LIME):** These methods assign an importance score to each input feature for a specific prediction (local explanation) or across the entire model (global explanation).
  - **Visual Explanations:** Techniques like saliency maps or Grad-CAM highlight specific parts of the input data (e.g., pixels in an image) that the model focused on when making a decision.
  - **Explanations by Simplification (Surrogate Models):** A simpler, interpretable model (e.g., a small decision tree) is trained to approximate the behavior of the complex black-box model, either globally or locally around a specific prediction.
  - **Explanations by Example:** Involves extracting representative examples or prototypes from the training data that are similar to the instance being predicted to help justify the outcome.
  - **Counterfactual Explanations:** These describe the smallest change to the input data that would alter the model's prediction (e.g., "If your income was $5,000 higher, your loan would have been approved").

3

Procedural and Design Approaches:

- **User-Centric Design:** Involves the end-users in the design process to ensure explanations meet their specific needs, knowledge levels, and context.
- **Integrate XAI Early:** Incorporate explainability into the AI development workflow from the outset, rather than as an afterthought.
- **Human-in-the-Loop:** Maintain meaningful human oversight so that a person can scrutinize the AI's recommendations, apply their own expertise, and be accountable for the final decision.
- **Auditing and Regulation:** Establish clear guidelines and regulatory frameworks that mandate specific levels of transparency and allow for independent audits of AI systems to ensure fairness and compliance.

Understand the difference between explainability and interpretability of algorithms

What are Interpretability and Explainability?

- **Interpretability:** refers to the ability to understand the decision-making process of an AI model. An interpretable model is t**ransparent in its operation and provides information about the relationships between inputs and outputs**. An interpretable algorithm can be explained clearly and understandably by a human being. Interpretability is therefore important to ensure that users can understand and trust artificial intelligence models.
- **Explainability:** pertains to the ability to explain the decision-making process of an AI model in terms understandable to the end user. An explainable model **provides a clear and intuitive explanation of the decisions made**, enabling users to understand why the model produced a particular result. In other words, explainability focuses on why an algorithm made a specific decision and how that decision can be justified.

Aspect	**Interpretability**	**Explainability**
Focus	Understanding the *model itself*	Understanding the *model's behavior or decisions*
Model Type	Usually applies to simple, transparent models (e.g., linear regression, decision tree)	Can be applied to complex "black-box" models (e.g., deep learning, ensembles)
Question Answered	"How does the model work?"	"Why did the model give this output?"
Approach	Direct transparency	Post-hoc (after training) explanations
Example	Interpreting weights in logistic regression	Using SHAP to explain a neural network's prediction

AI/ML algorithmic robustness(adversarial attacks,minimizing security risks)

Algorithmic robustness means how **stable and reliable** an AI/ML model is when faced with **unexpected, noisy, or maliciously altered inputs**.

In simple words:

> A robust model continues to perform well even when the input data slightly changes or contains attacks/errors.

Robustness ensures that the model's behavior remains **consistent, trustworthy, and safe** under real-world conditions.

🔹 2. Adversarial Attacks

Adversarial attacks are **intentional manipulations** of input data designed to **fool AI models** into making wrong predictions — even though changes may be **imperceptible to humans**.

These are small, carefully crafted perturbations that exploit the model's vulnerabilities.

🔸 Types of Adversarial Attacks

Type	Description	Example
Evasion Attack	Modify input data at prediction time to mislead the model.	Adding small noise to an image so that a "stop sign" is classified as a "speed limit sign."
Poisoning Attack	Tamper with the training data to corrupt the model.	Injecting mislabeled samples into training data so the model learns wrong patterns.
Model Inversion Attack	Try to extract private or sensitive information from the model's outputs.	Guessing details of individuals in a medical dataset used to train an AI.
Membership Inference Attack	Determine whether specific data points were used in training.	Identifying if a person's data was part of the training set, violating privacy.

🔸 Example (Evasion Attack on Image Model)

An attacker adds a nearly invisible noise pattern to a picture of a **"stop sign."**

- To humans → it still looks like a stop sign.
- To the AI → it's now recognized as a **"speed limit" sign.**

🚗 This is dangerous for autonomous vehicles or security systems.

🔹 3. Minimizing Security Risks (Improving Robustness)

To make models **robust against adversarial and security threats**, several strategies are used:

(a) Adversarial Training

- Train the model on both **normal** and **adversarially perturbed** samples.
- The model learns to recognize and resist manipulated inputs.

🧠 *Idea:* "If I've seen such attacks before, I won't be fooled again."

(b) Defensive Distillation

- Train the model to **smooth its decision boundaries**, making it less sensitive to small input changes.

(c) Input Preprocessing

- Apply filters, noise reduction, or normalization to remove adversarial noise before feeding data to the model.

(d) Model Monitoring & Anomaly Detection

- Continuously monitor model inputs and outputs for **unusual patterns** that might indicate an attack.

(e) Secure Data and Training Pipelines

- Ensure **data integrity** (no tampering during collection or training).
- Use encryption, access control, and secure APIs.

(f) Regular Audits and Testing

- Conduct **robustness testing** under simulated adversarial conditions.
- Evaluate model performance not only on accuracy but also **resilience** to perturbations.

🔹 4. Why Robustness Matters

- **Safety:** Prevents AI failures in critical systems (healthcare, self-driving cars, finance).
- **Security:** Protects against malicious manipulation.
- **Trust:** Builds user confidence in AI decisions.
- **Fairness:** Prevents attackers from exploiting vulnerabilities to target specific groups.