Introduction to Responsible AI

Course Code: MDM5

B. Tech : VII Semester

Prepared by:

Mr. Ashish Kakne

Assistant Professor

Department of Information Technology

G. H. Raisoni College of Engineering and Management,

Nagpur

UNIT I | 6 HOURS

Introduction to Responsible AI & Foundations

LECTURE 1

What is Responsible AI?

Responsible AI focuses on building systems that are safe, ethical, and trustworthy, instead of just aiming for mathematical accuracy.

- **Human-First Tech:** Making sure automation respects societal values and doesn't discriminate.
- **Error Control:** Understanding what occurs in that critical 1% failure boundary.

🇮🇳 Nagpur Ration Card Example:

If an automated biometric fingerprint reader rejects a family's biometric match, a responsible backup process must ensure they still receive their monthly food grains without delay.

LECTURE 1 VISUAL

Responsible AI Framework



LECTURE 2

The Background and Evolution of AI

Unchecked rapid growth of ML algorithms has transformed our world. We must trace the journey from simple rules to highly complex, opaque models.

Rule-Based Era

Simple "If-Then" programs. Highly predictable but unable to handle dynamic real-world scenarios.

Machine Learning

Algorithms discovering patterns directly from historical datasets without human coding intervention.

Deep Neural Nets

Opaque models mimicking human brains. Exceptionally accurate but completely non-transparent.

LECTURE 3

Autonomy in AI Systems

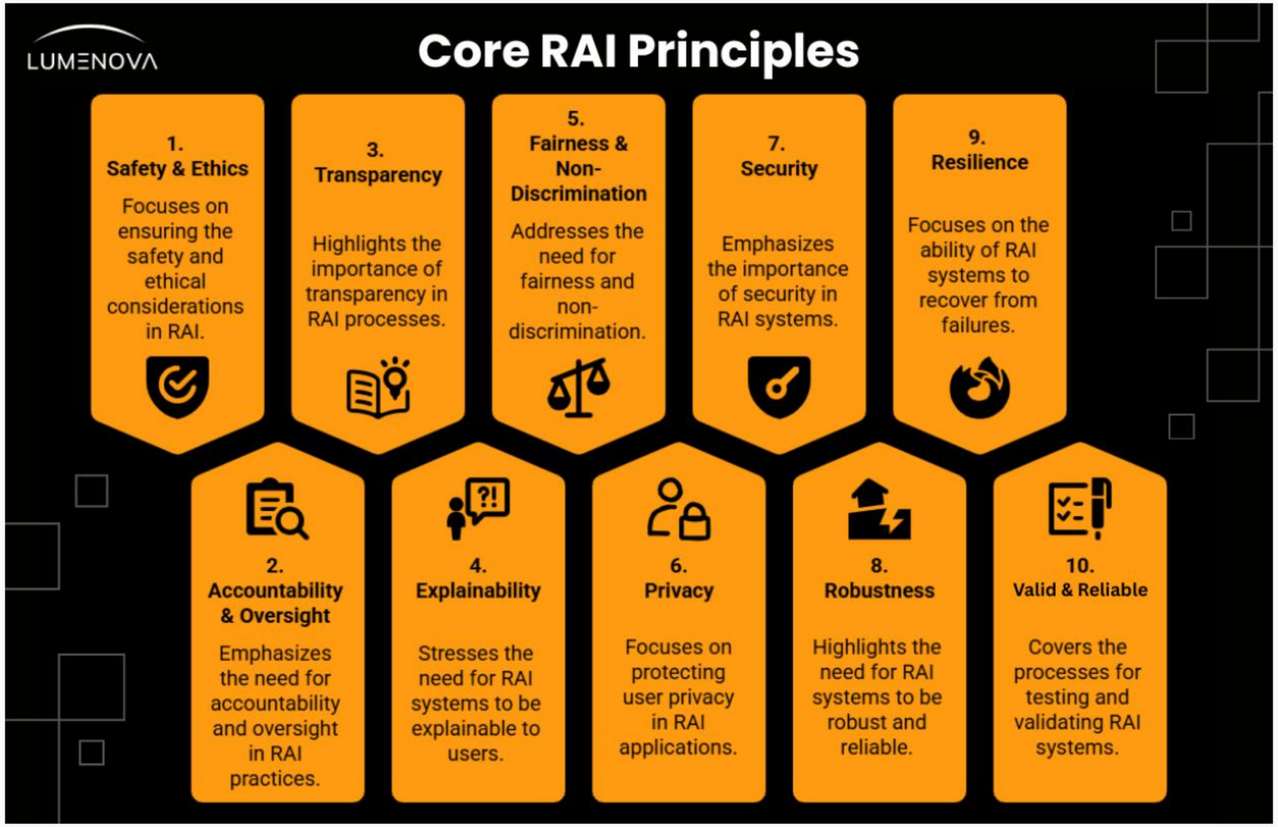
Autonomy refers to the level of independence an AI system possesses to take actions and execute choices without real-time human oversight.

- **Full Autonomy:** Fully driverless navigation on crowded roads.
- **Partial Autonomy:** An automatic emergency braking assist tool in modern passenger vehicles.

> 🇮🇳 **Local Nagpur Traffic Context:**
> An automated red-light speed camera system that issues fines independently must operate with high-precision validation logic to prevent wrong penalties.

LECTURE 3 VISUAL

Autonomous Mapping and Navigation



LECTURE 4

Adaptability: Learning from Trends

Adaptability measures how dynamically an AI system alters its parameters when exposed to updated trends and shifting live conditions.

- **Continuous Learning:** Recalibrating forecast models dynamically as consumer buying choices change.
- **Feedback Vulnerability:** If fed with malicious, manipulated user inputs, the system is highly prone to picking up wrong patterns.

> 🇮🇳 **Indian Financial Market Context:**
> An automated algorithmic trader on the National Stock Exchange (NSE) must adjust dynamically during volatile budget announcements to avoid executing high-loss panic trades.

LECTURE 5

Interaction: Humans & AI Collaboration

Designing user interfaces that keep humans in control while letting AI do the complex data processing.

Human-In-The-Loop

AI scans and filters options, but the critical final decision is executed by a qualified human expert.

Local Language UI

Allowing regional communication (like Marathi and Hindi speech) to assist rural farmers and merchants.

Bilateral Trust

Ensuring users interact with clear, safe systems without worrying about background personal data leaks.

LECTURE 6

Responsible Generative AI & Applications

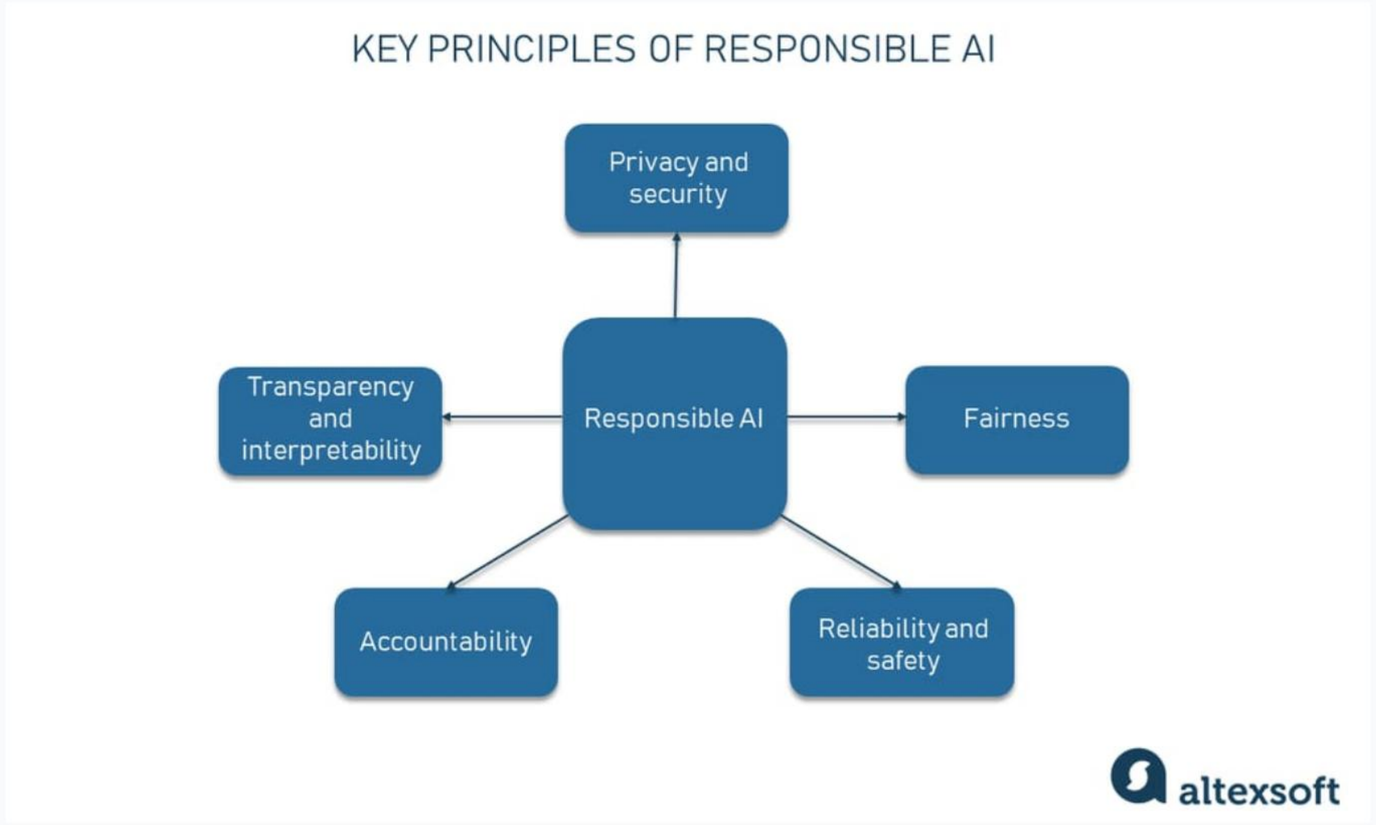
Generative models (like LLMs creating regional text and imagery) raise distinct trust and safety concerns.

- **Hallucinations:** The tendency of LLMs to generate highly convincing but totally incorrect facts.
- **Media Misuse:** Deepfakes spreading fake rumors about community figures or political events.

> 🇮🇳 **Government Service Integration:**
> An Indian automated tax query responder must base its answers strictly on official laws, completely avoiding guessing.

LECTURE 6 VISUAL

Generative AI Governance Framework



UNIT II | 10 HOURS

Introduction to Explainable AI (XAI) & Tools

LECTURES 7 & 8

XAI: Solving the "Black Box" Problem

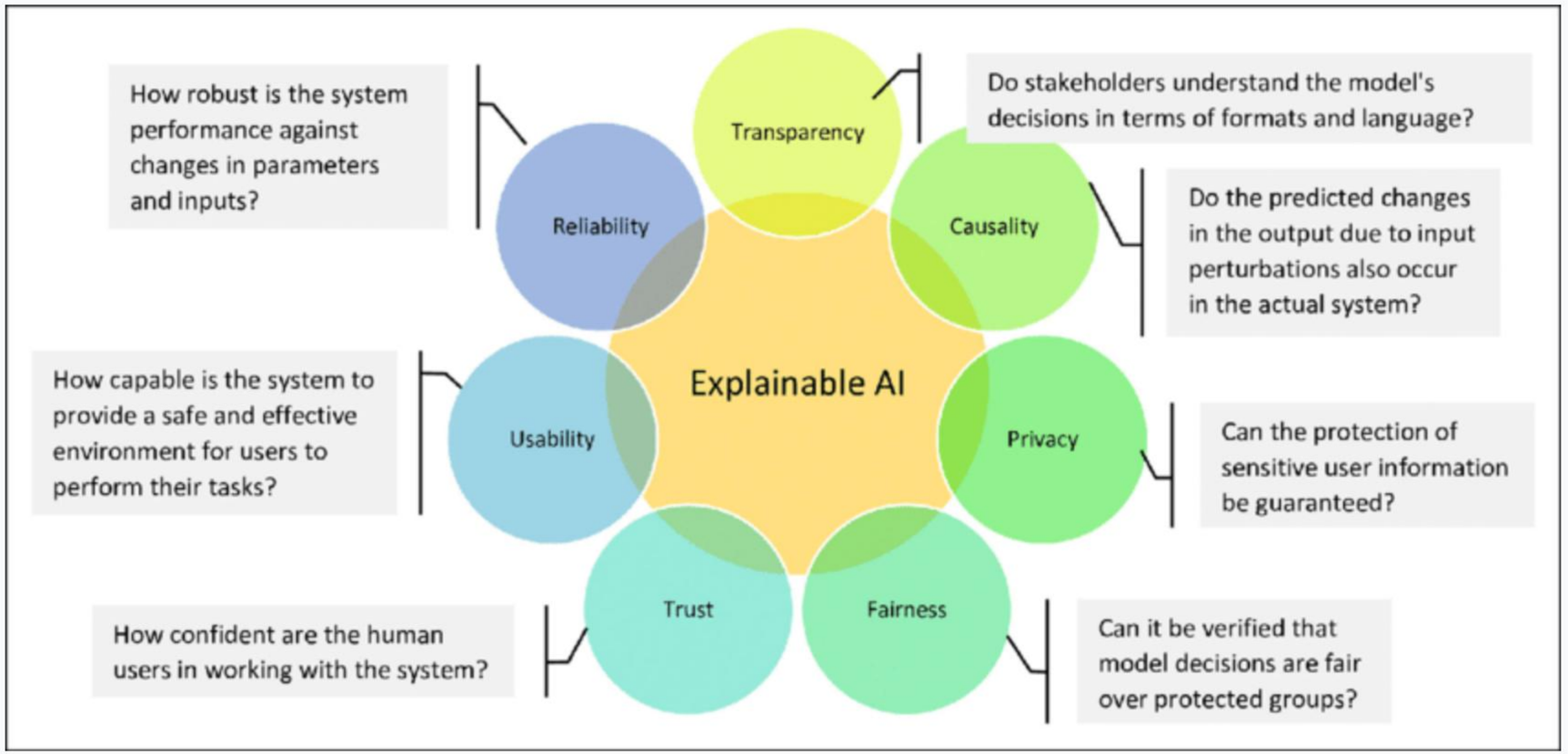
Deep neural networks make complex predictions, but cannot explain the internal reasoning behind their output. XAI tools bridge this gap.

- **Right to Explanation:** Users have a right to know the parameters behind automated choices affecting them.
- **Debugging Models:** Developers must pinpoint exactly why a model failed in production.

> 🇮🇳 **Medical Diagnosis Context:**
> If an AI system spots a tumor in an X-ray at a local Nagpur hospital, the doctor needs to see highlighted pixels showing why it reached that conclusion.

LECTURES 7 & 8 VISUAL

Explainable AI (XAI) Architecture



LECTURE 9

Transparency vs Interpretability

These two terms are fundamental to XAI but have distinct engineering goals.

Transparency

Providing full access to details of the model, data, weights, and parameters. It answers: *"Can we see the recipe?"*

Interpretability

The ease with which a human can predict the model's output changes when inputs are varied. It answers: *"Why this result?"*

The Trade-Off

Simple models are highly interpretable but have lower accuracy. Complex deep models are highly accurate but hard to interpret.

LECTURE 10

Models for Interpreting AI Systems

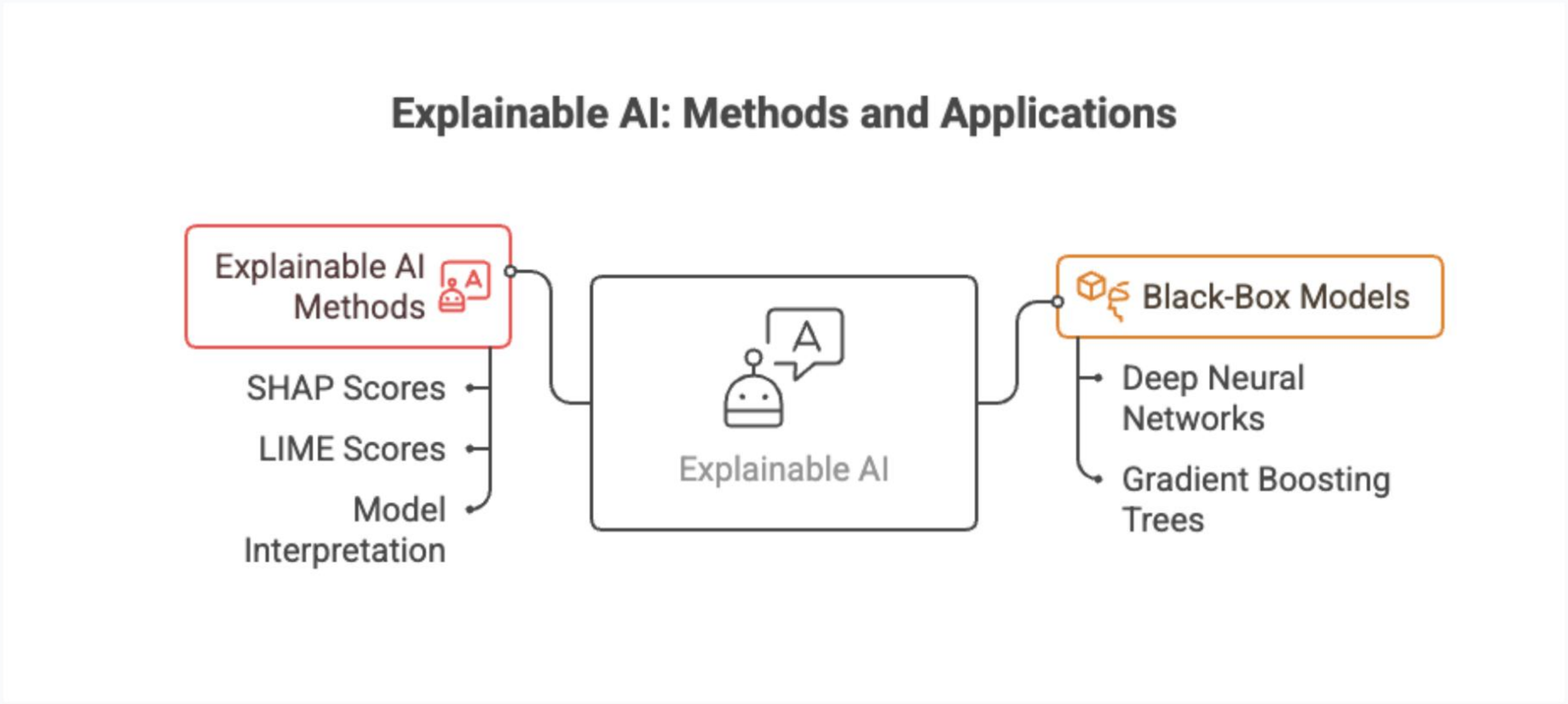
We classify models into two broad interpretation approaches based on their internal structure.

- **Intrinsic Models:** Models that are simple enough to be interpretable by nature (e.g., Linear Regression, Decision Trees).
- **Post-Hoc Models:** Separate interpretation methods applied to complex models *after* they have finished training.

> **🇮🇳 Crop Yield Prediction Example:**
> A simple formula calculating yield based on rainfall is intrinsic. A deep vision model assessing leaf health needs post-hoc heatmaps to explain its choices.

LECTURE 10 VISUAL

Intrinsic vs Post-Hoc Interpretability



LECTURE 11

Explaining Decisions of AI Systems

To establish human trust, automated systems must explain why they chose "Option A" over "Option B".

- **Contrastive Explanations:** Detailing what changes would convert a negative decision into a positive one.
- **Global vs Local:** "Global" explains the overall system logic, while "Local" explains a specific individual outcome.

> 🇮🇳 **Home Loan Approval Context:**
> Instead of just saying "Loan Rejected," the system should explain: "If your monthly income was ₹10,000 higher, your application would be approved."

LECTURE 12

Evaluating Interpretability and Explanation

How do we measure if an XAI explanation is mathematically accurate and useful to humans?

Fidelity Score

Measures how closely the proxy model matches the true black-box decision logic.

Human Evaluation

Testing real users to see if the explanation helps them complete tasks faster and with higher confidence.

Robustness

Ensuring small changes in inputs do not lead to wildly erratic explanations.

LECTURE 13

Tools of Explainable AI: LIME

LIME Framework

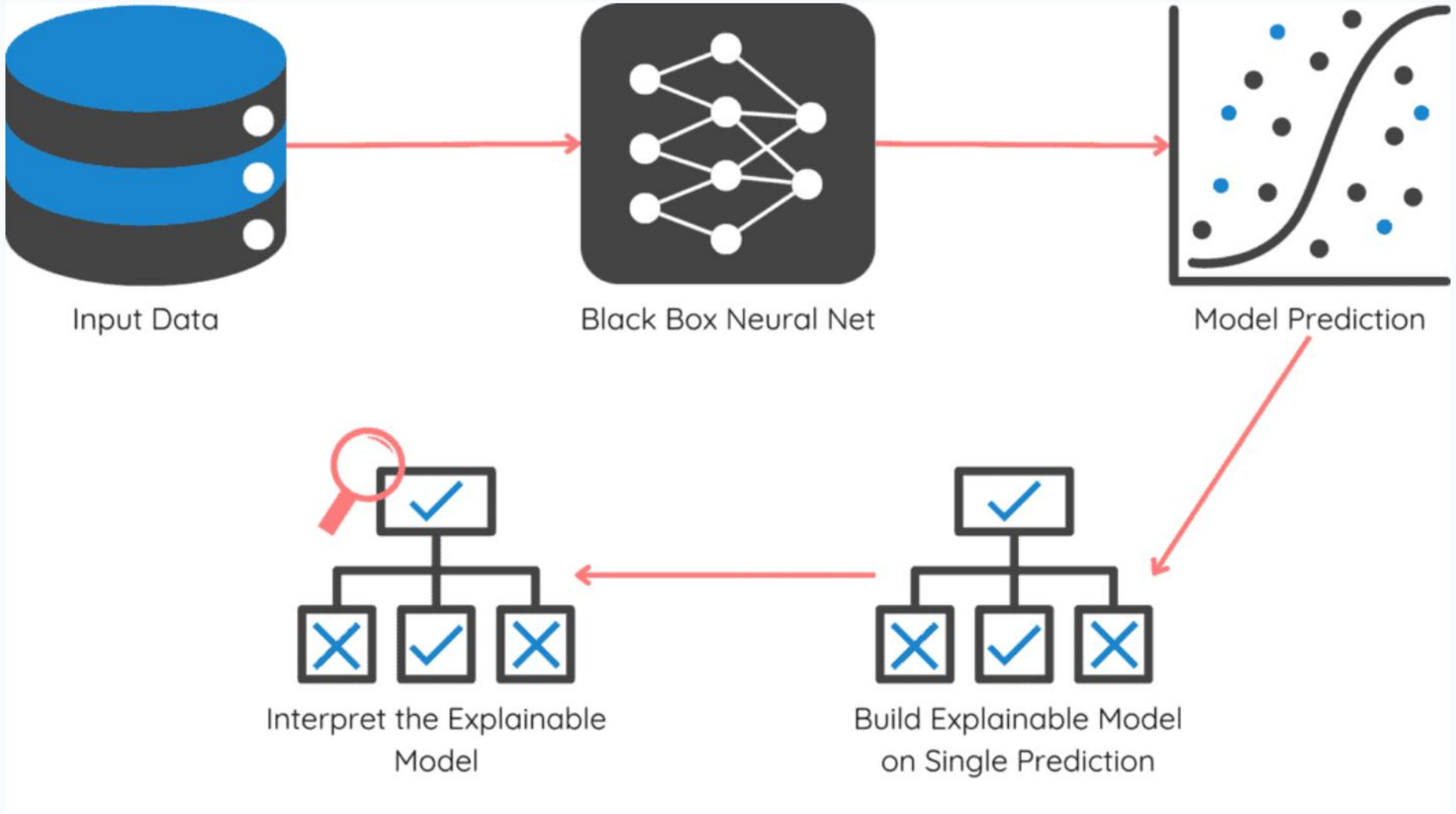
Local **I**nterpretable **M**odel-Agnostic **E**xplanations.

- **How it works:** It slightly perturbs the inputs around a single data point and tracks how the outputs shift.
- It fits a simple, interpretable linear model locally to explain that specific decision.

> 🇮🇳 **Crop Insurance Example:**
> Explaining why a specific farmer's claim was denied by analyzing local weather and crop metrics.

LECTURE 13 VISUAL

LIME Attribution Visualization



LECTURE 14

Tools of Explainable AI: SHAP

SHAP Framework

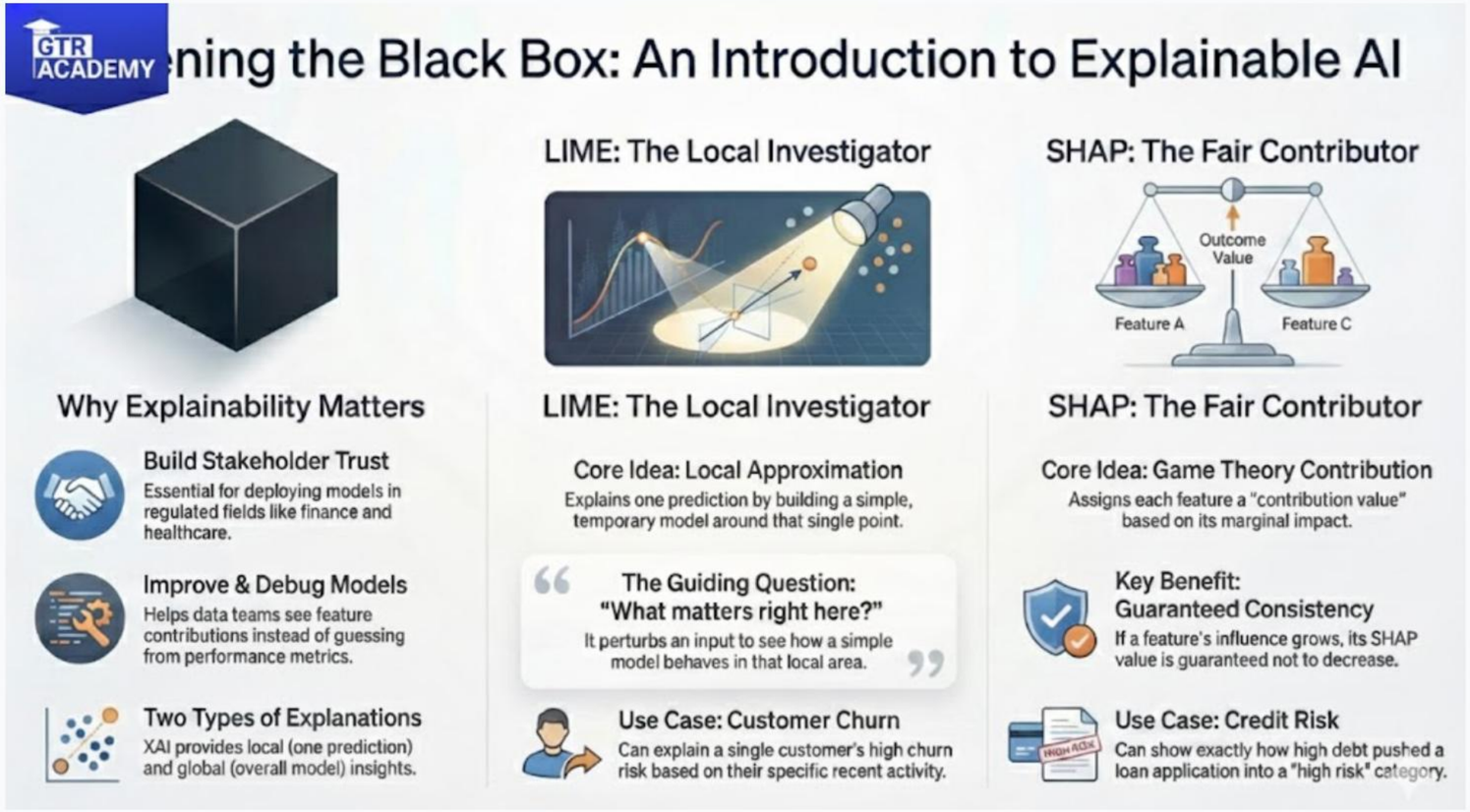
SHapley **A**dditive ex**P**lanations.

- **Theoretical Basis:** Built on Game Theory concepts.
- Treats each input feature as a player in a game, measuring exactly how much credit or blame they deserve for the final score.

> 🇮🇳 **Indian Credit Scoring Context:**
> Determining exactly how much weight salary, past loans, and age played in assessing an applicant's credit risk.

LECTURE 14 VISUAL

SHAP Feature Impact Summary



LECTURE 15

Responsible Research & Innovation (RRI)

RRI ensures that scientific research and technological progress align with the values, needs, and expectations of society.

- **Anticipation:** Predicting potential negative impacts before developing technologies.
- **Deliberation:** Discussing projects with diverse public groups, not just engineering experts.
- **Responsiveness:** Adapting research paths when social or ethical concerns are flagged.

> 🇮🇳 **Public Policy Focus:**
> When planning large-scale smart public transport networks, cities should actively gather feedback from local commuters.

LECTURE 16

The ART of AI Framework

A structured approach to designing ethical AI systems around three pillars.



Accountability

Identifying who is responsible for the decisions and impacts of an AI system (the developer, the owner, or the user).



Responsibility

Ensuring AI systems align with human rights, social values, and local regulations.



Transparency

Keeping research details, data collection, and model performance metrics open and inspectable.

UNIT III | 10 HOURS

Ethics in Responsible AI & Moral Agents

LECTURES 17 & 18

Can AI Systems Be Ethical?

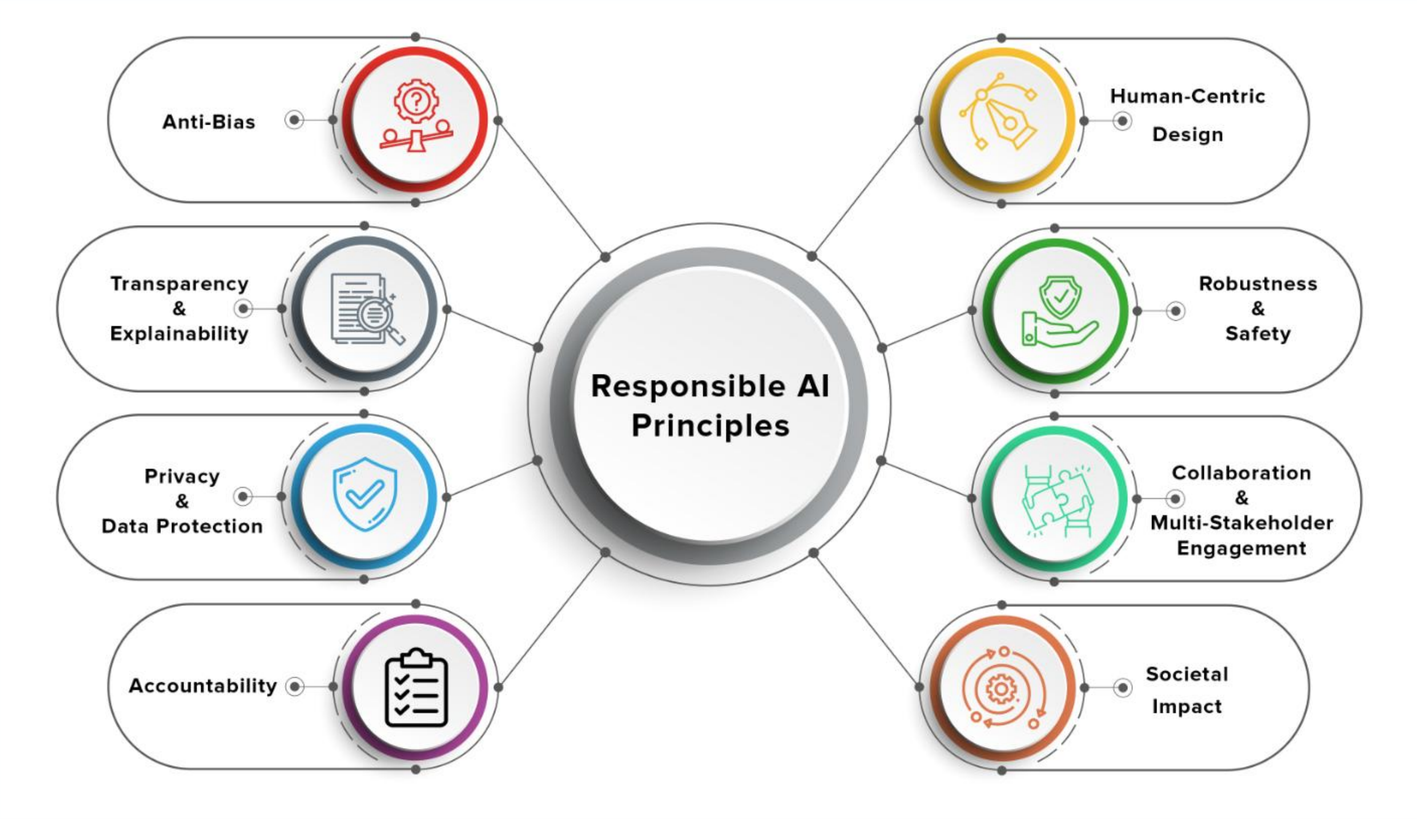
Ethics is about "Right vs. Wrong". Can a machine have a conscience, or does it simply mimic human patterns?

- AI doesn't have feelings, but it follows the rules, objectives, and parameters we code into it.
- If we train models on biased data, they will display unethical behaviors.

> 🇮🇳 **Healthcare Resource Allocation:**
> If an algorithm must allocate limited hospital beds during a health crisis, how should it fairly rank patients?

LECTURES 17 & 18 VISUAL

Ethical Tech Balance



LECTURE 19

Introduction to Ethical Action

An ethical action is one that is morally acceptable and aligned with standard values of societal good.

- **Algorithmic Action:** Translating values into computable constraints.
- **Values Alignment:** Ensuring software parameters map directly to human safety.

> 🇮🇳 **E-Commerce Context:**
> Ensuring delivery algorithms in Nagpur do not over-work delivery executives beyond legal hours for micro-efficiency gains.

LECTURE 20

Ethical Reasoning Approaches by AI

How do we program ethical decision-making into machines? We study three traditional philosophical approaches.

Utilitarianism

Aiming for the greatest good for the greatest number of people (e.g., routing traffic to minimize city-wide pollution).

Deontology

Strictly following pre-defined rules and duties, regardless of the outcome (e.g., "Never lie, never cheat").

Virtue Ethics

Acting in alignment with character virtues like wisdom, fairness, and compassion.

LECTURE 21

Designing Artificial Moral Agents

Artificial Moral Agents (AMAs) are systems designed to evaluate their actions and make choices based on ethical principles.

- **Top-Down Design:** Embedding logical rules directly into the code (e.g., Asimov's Laws of Robotics).
- **Bottom-Up Design:** Training the model to learn ethical behavior by observing positive human interactions.
- **Hybrid Approaches:** Combining foundational rules with ongoing observation and learning.

> 🇮🇳 **Indian Values Context:**
> Teaching customer support bots to speak respectfully to elders, reflecting traditional family values.

LECTURE 22

Ensuring Responsible AI in Practice

Moving from philosophy to direct software engineering practices.

- **Impact Assessments:** Mandatory ethical evaluations before software development begins.
- **Testing & Validation:** Running stress tests to observe how models react to unexpected situations.
- **Continuous Monitoring:** Spotting and correcting bias or errors in real-time as the AI operates.

> 🇮🇳 **Tech Park Practice:**
> Engineers in Indian IT hubs must complete mandatory model audit checklists before publishing any model to public clouds.

LECTURE 23

Governance for Responsible AI

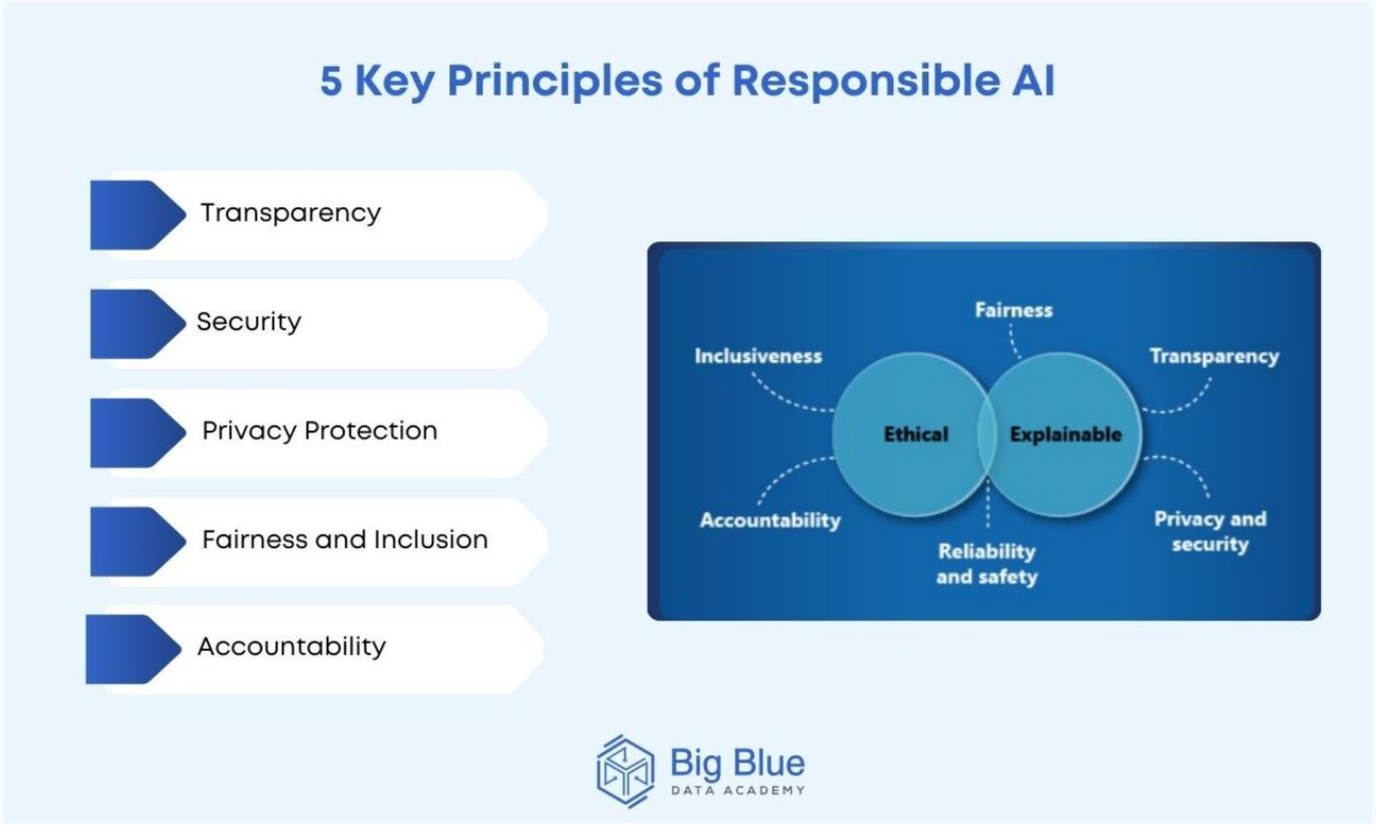
We need policies, regulations, and institutional frameworks to enforce accountability in AI development.

- **Global Standards:** Frameworks like the EU GDPR and UNESCO AI Guidelines.
- **National Frameworks:** India's NITI Aayog guidelines on "Responsible AI for All."

> 🇮🇳 **Government Services:**
> Ensuring Indian public database integrations protect citizen privacy while improving automated service delivery.

LECTURE 23 VISUAL

AI Governance Standards



LECTURE 24

Codes of Conduct

Establishing practical, everyday standards for development teams inside organizations.

- **Voluntary Codes:** Ethical codes of conduct adopted by organizations (e.g., "Don't use manipulative UI design").
- **Mandatory Standards:** Audits required for safety-critical systems like aviation, automotive, and healthcare.
- **Whistleblower Protection:** Enabling engineers to report unethical data practices safely.

> 🇮🇳 **Developer Focus:**
> Ensuring software engineers are trained to check datasets for representative bias before starting development.

LECTURE 25

Inclusion and Diversity

Inclusion means building AI systems that work reliably for everyone, regardless of background, language, or physical ability.

Representative Data

Ensuring datasets include faces, voices, and dialects from diverse regions and demographics.

Accessibility

Designing interfaces that work for users with visual, auditory, or motor challenges.

Language Diversity

Supporting local languages and dialects to make information accessible to rural communities.

LECTURE 26

The AI Narrative & Society

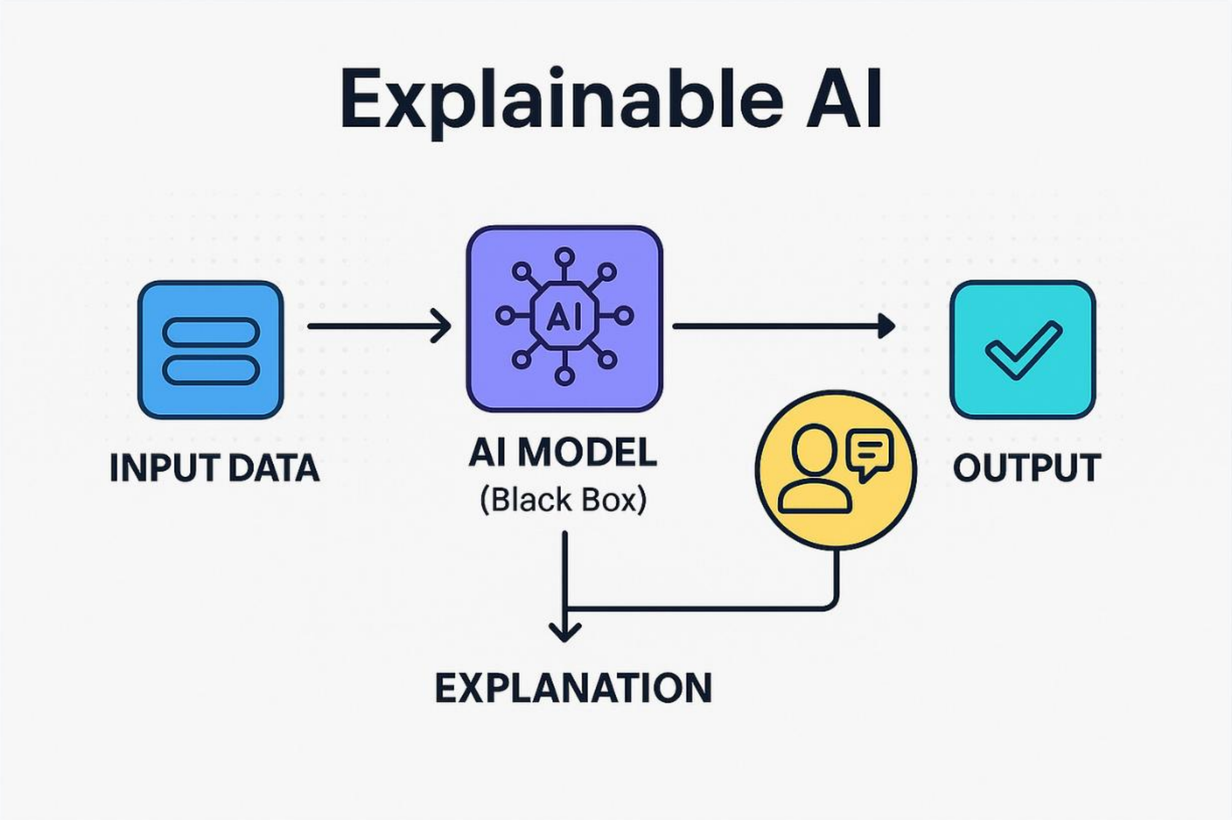
Media and pop culture often shape our view of AI, creating a gap between fictional worries and real-world issues.

- **The Sci-Fi Fear:** Worrying about sentient robots taking over the world.
- **The Real Danger:** Facing daily issues like bias, privacy loss, and algorithmic discrimination.

> 🇮🇳 **Job Market Focus:**
> The real challenge isn't robots taking all jobs, but upskilling local workers to thrive alongside new automated tools.

LECTURE 26 VISUAL

Real Societal Impacts of AI



LECTURES 27 & 28

The Ethical Status of AI Systems

As AI tools become more advanced and conversational, we must consider how we treat them and define their legal standing.

- **Legal Personhood:** Should highly advanced AI systems hold legal rights, or are they strictly tools of property?
- **Moral Standing:** Does mistreating a conversational AI agent have an ethical impact on our own behavior?
- **Accountability Gap:** If an autonomous system causes financial harm, does the liability rest on the user or the developer?

> 🇮🇳 **Responsibility Focus:**
> Understanding that from a legal perspective, final accountability always remains with the human creators and operators.

UNIT IV | 9 HOURS

Bias & Fairness in Artificial Intelligence

LECTURES 29 & 30

Bias and its Types

Algorithmic bias occurs when a system systematically outputs unfair or inaccurate results for specific groups of people.

- **Historical Bias:** Training data that reflects old societal inequalities (e.g., older hiring datasets with fewer women in tech).
- **Representation Bias:** Underrepresenting certain demographics in the training dataset.

> 🇮🇳 **Regional Language Context:**
> A speech model trained only on urban speakers failing to understand rural regional accents.

LECTURES 29 & 30 VISUAL

Bias & Fairness Balance



LECTURE 31

Fairness Concepts & Models

How do we define fairness mathematically? There are multiple, sometimes conflicting, ways to measure it.

Demographic Parity

Ensuring the selection or approval rate is equal across all demographic groups, regardless of characteristics.

Equal Opportunity

Ensuring the true positive rate is equal across all groups (qualified candidates have the same chance of approval).

Individual Fairness

Ensuring similar individuals receive similar outcomes, regardless of their demographic group.

LECTURES 32 & 33

Metrics for Bias & Quantifying Fairness

Using statistical metrics to identify and measure bias in prediction models.

- **Disparate Impact Ratio:** Comparing the selection rate of a protected group to that of a control group.
- **The 80% Rule:** A selection rate for any group that is less than 80% of the rate for the highest group is considered evidence of adverse impact.

🇮🇳 Indian College Scholarship Analogy:

If 100 male applicants but only 40 female applicants from the same eligibility bracket are selected by an AI, the ratio is 40%—highlighting systemic bias.

LECTURES 34 & 35

COMPAS Case Study

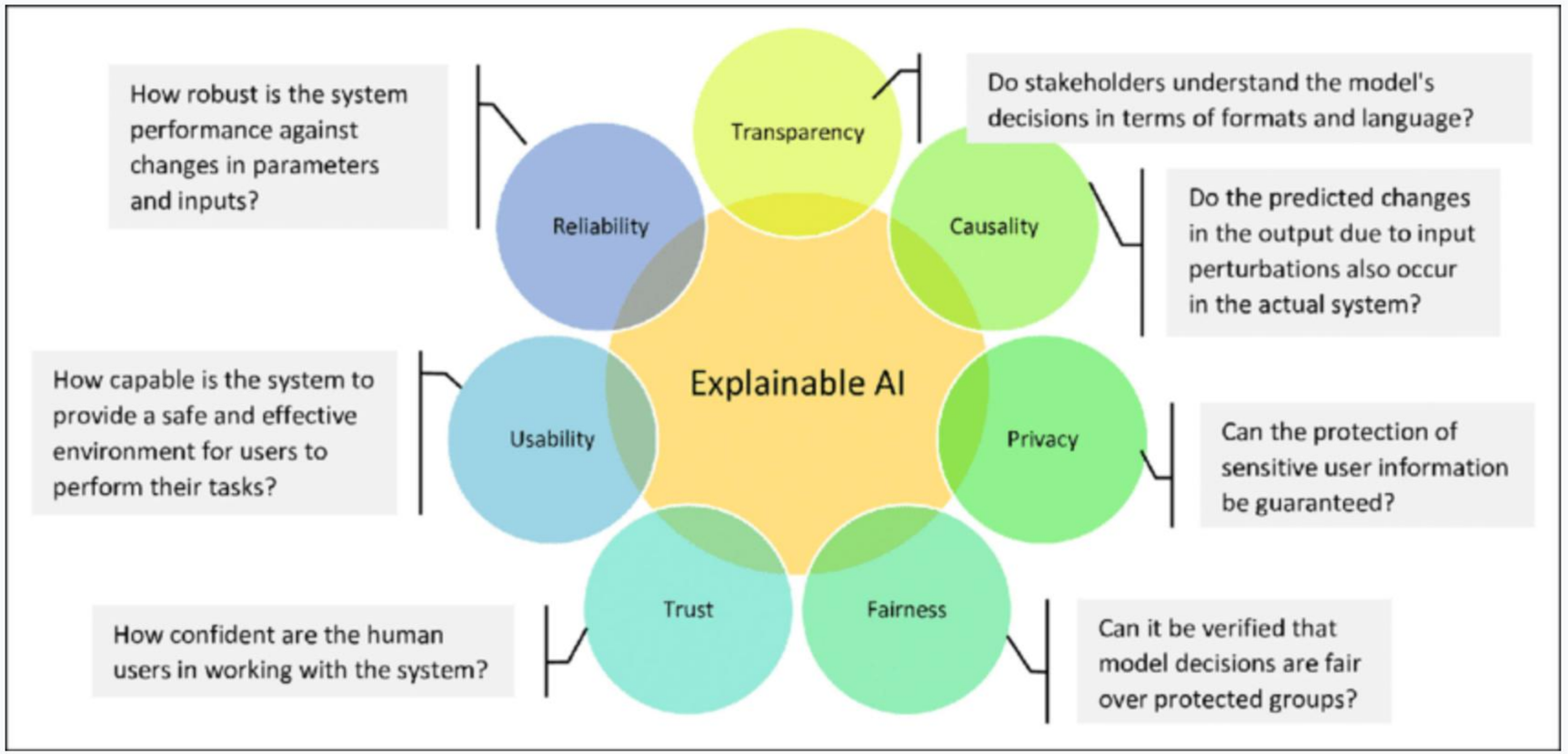
The COMPAS software, used in courts to assess criminal recidivism, revealed how biases can be built directly into risk-scoring algorithms.

- **The Bias:** The algorithm falsely flagged black defendants as high risk twice as often as white defendants.
- **The Impact:** Highlighting the danger of using historical data that contains systemic biases.

> **Critical Lesson:**
> We must audit and test algorithms for fairness before deploying them in critical public decision systems.

LECTURES 34 & 35 VISUAL

Recidivism Risk Score Bias



LECTURES 36 & 37

Code Walkthrough of Bias Matrices

A practical walkthrough showing how to identify and measure bias using open-source Python toolkits like AIF360.

Python Reference Code:

```
from aif360.metrics import BinaryLabelDatasetMetric
# Compute disparate impact ratio
metric = BinaryLabelDatasetMetric(dataset, unprivileged_groups, privileged_groups)
print("Disparate Impact:", metric.disparate_impact())
```

- **Pre-processing:** Reweighting the dataset before training to balance representation.
- **In-processing:** Adding fairness constraints directly into the model's loss function.
- **Post-processing:** Adjusting decision thresholds after training to ensure balanced outcomes.

UNIT V | 9 HOURS

Privacy Preserving AI Techniques & Security

LECTURES 38 & 39

Privacy in AI Systems

AI models require large amounts of data, which raises important concerns about user privacy and data security.

- **Data Leaks:** The risk of models exposing personal user information from their training data.
- **Regulatory Compliance:** Aligning data collection with frameworks like GDPR and local privacy laws.

> 🇮🇳 **Digital Personal Data Protection Act (DPDP):**
> India's regulations requiring clear user consent and strict safeguards for all personal data.

LECTURES 38 & 39 VISUAL

Digital Cyber Security Shield



LECTURES 40 & 41

Securing Federated Learning

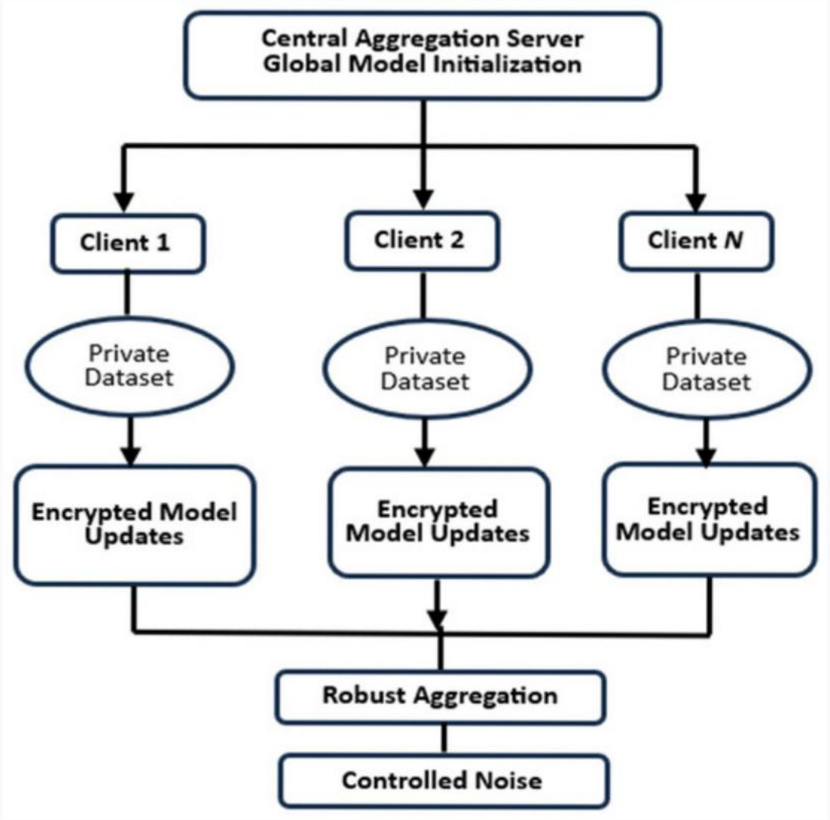
Federated Learning allows models to train on decentralized edge devices without centralizing raw user data.

- **Data Minimization:** Raw user files stay on the local device, protecting privacy.
- **Aggregation:** Only encrypted model updates are sent to the central server to improve the global model.

> 🇮🇳 **Smart Keyboard Example:**
> Your mobile keyboard learns your typing habits and local slang locally, without sending your private messages to a central server.

LECTURES 40 & 41 VISUAL

Decentralized Federated Nodes



LECTURE 42

Differential Privacy (DP)

Differential Privacy is a mathematical framework that protects individual identities in a dataset by adding calculated noise.

Adding Noise

Injecting random mathematical noise to hide individual data points while keeping overall trends accurate.

Aggregated Analysis

Allowing researchers to draw insights from large datasets without identifying specific individuals.

Privacy Budget

Balancing the amount of noise added to maintain both data privacy and utility.

LECTURES 43 & 44

Homomorphic Encryption

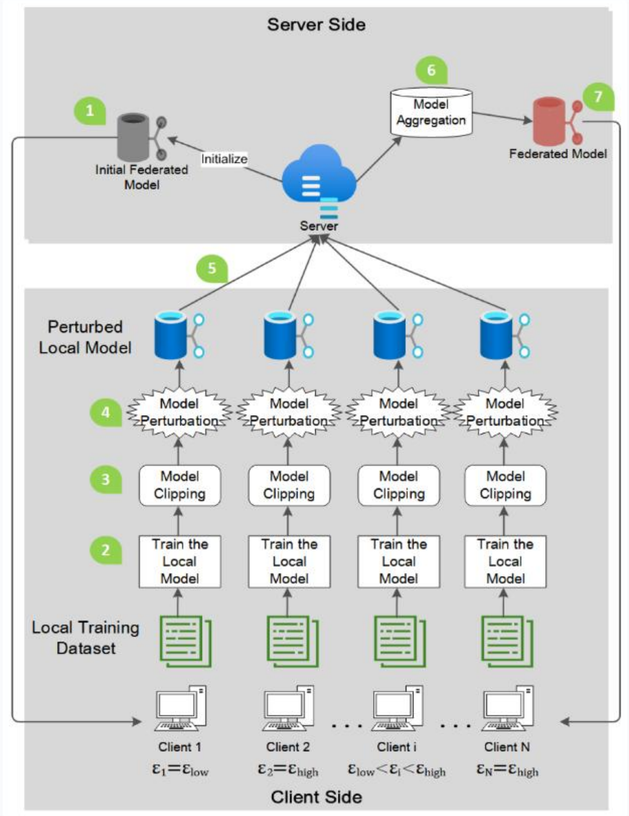
Homomorphic Encryption allows mathematical calculations to be performed directly on encrypted data without decrypting it first.

- **Secure Cloud Compute:** Organizations can send encrypted data securely to third-party cloud servers for modeling.
- **End-to-End Security:** Keeps the information secure throughout its lifecycle.

> 🇮🇳 **Nagpur Healthcare Network:**
> Hospitals can analyze encrypted patient records securely on public cloud servers without exposing sensitive medical histories.

LECTURES 43 & 44 VISUAL

Homomorphic Process flow



Course Summary: Main Takeaways

Unit Focus	Core Concept & Skill Acquired	Why it Matters in Practice
Unit I & II (Explainability)	Using tools like LIME and SHAP to explain how complex models make decisions.	Builds user trust and simplifies system debugging.
Unit III & IV (Ethics & Bias)	Identifying and mitigating bias using fairness metrics.	Prevents algorithmic discrimination against underrepresented groups.
Unit V (Privacy & Security)	Using techniques like Federated Learning and Homomorphic Encryption to secure data.	Protects citizen privacy and ensures regulatory compliance.

G. H. RAISONI COLLEGE OF ENGINEERING AND MANAGEMENT, NAGPUR

Questions & Answers

Let's discuss how we can build safe and ethical AI systems together.

Thank You!

Department of Information Technology | Semester VII | Nagpur

Image Sources



https://verasolutions.org/wp-content/uploads/2025/04/Principles-graphic-edit-2-1024x536.jpg

Source: verasolutions.org



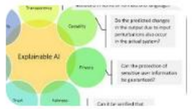
https://www.lumenova.ai/wp-content/uploads/principles-of-responsible-ai-1.png

Source: www.lumenova.ai



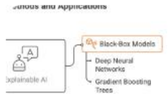
https://www.altexsoft.com/static/blog-post/2023/11/6d2dceff-ede6-42d6-9d97-38715e12046f.jpg

Source: www.altexsoft.com



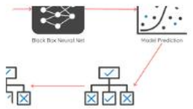
https://miro.medium.com/1*MzchgMabf5RB0LIvCH6SMg.png

Source: vtantravahi.medium.com



https://miro.medium.com/1*538KJsMPKCoWubdlVLeLoQ.png

Source: medium.com



https://viso.ai/wp-content/uploads/2024/02/Working-of-LIME-method-1060x596.png

Source: viso.ai

Image Sources



https://gtracademy.org/wp-content/uploads/2026/01/XAI_Cretive_logo.png

Source: gtracademy.org



https://insights.daffodilsw.com/hs-fs/hubfs/Allen/Responsible%20AI.png?width=3300&height=1908&name=Responsible%20AI.png

Source: insights.daffodilsw.com



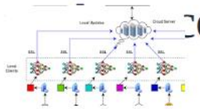
https://bigblue.academy/uploads/images/blog/responsible-ai/principles-of-responsible-ai.jpg

Source: bigblue.academy



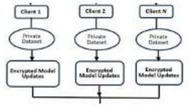
https://miro.medium.com/v2/resize:fit:1400/1*EuW6ZFK0Qc4uj8h3gBwT-A.png

Source: medium.com



https://www.mdpi.com/mathematics/mathematics-12-03194/article_deploy/html/images/mathematics-12-03194-g001.png

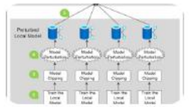
Source: www.mdpi.com



https://www.mdpi.com/applsci/applsci-15-06878/article_deploy/html/images/applsci-15-06878-g001-550.jpg

Source: www.mdpi.com

Image Sources



https://www.mdpi.com/applsci/applsci-13-04168/article_deploy/html/images/applsci-13-04168-g001.png

Source: www.mdpi.com