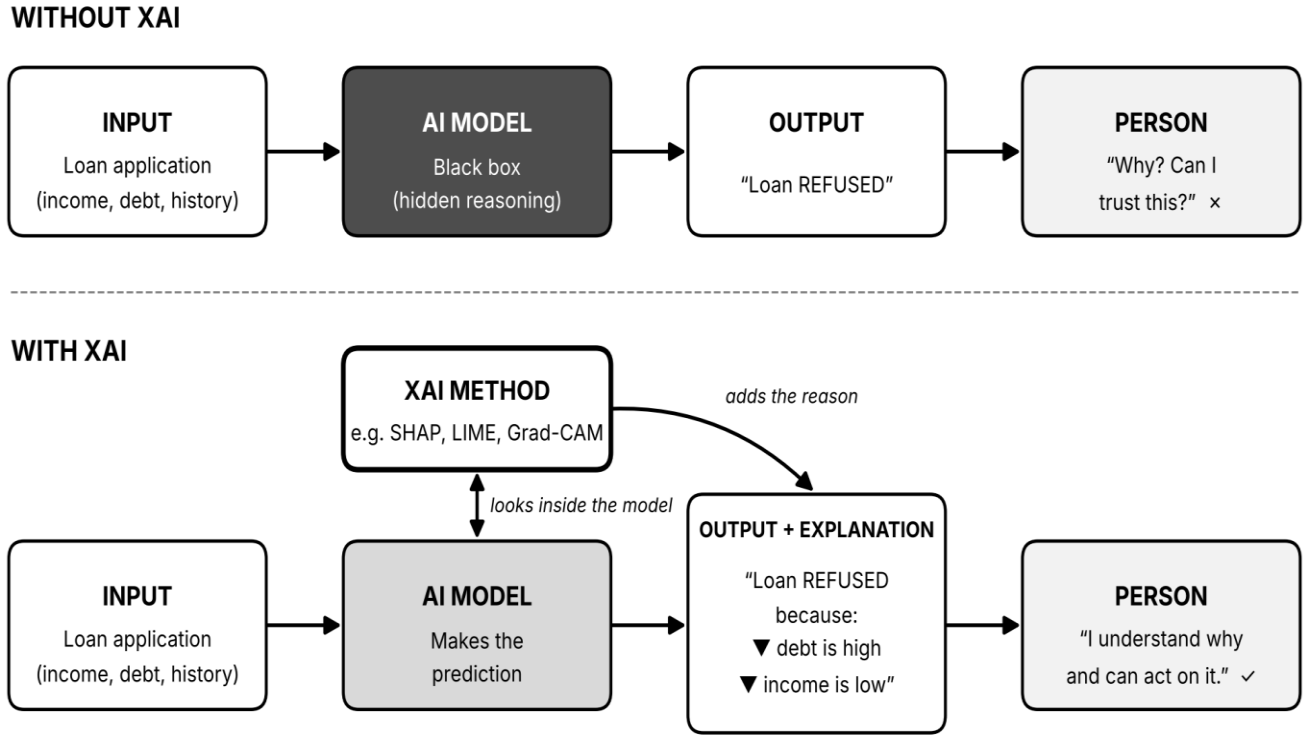
Explainable AI (XAI): A Compact Tutorial

Concepts · Methods · Tools · Applications · Regulation · Landmark Research | ***@SSRoy***

1. What Is Explainable AI and Why Does It Matter?

Many powerful AI models (deep neural networks, large language models) are **black boxes**: they give an answer but not the reason. **Explainable AI (XAI)** is the set of methods that make an AI system's decisions understandable to people. A doctor should know *why* a model flags an X-ray; a bank customer should know *why* a loan was refused. XAI builds **trust**, helps engineers **debug** models, exposes **bias**, supports **accountability**, and is now demanded by **law** (Section 7).

Simple analogy: a black-box model is a student who writes only the final answer; an explainable model "shows the working", so a teacher can check whether the answer is right for the right reason.



XAI turns a black-box answer into an answer with reasons that people can check, trust and act on.

2. Essential Terminology

Term	Meaning in simple words
Black-box vs glass-box model	Black-box: logic hidden or too complex (deep nets, large ensembles). Glass-box (white-box): logic readable by a human (small tree, linear model).
Interpretability vs explainability	Interpretability: how easily a human understands the model itself. Explainability: giving a human-understandable reason for a decision, often via an extra method.
Transparency	Openness about data, design, training and limits of a system (e.g. model cards, datasheets for datasets).
Intrinsic (ante-hoc) vs post-hoc	Intrinsic: the model is interpretable by design. Post-hoc: a separate method explains an already-trained model.
Global vs local explanation	Global: how the model behaves overall (which features matter in general). Local: why this one prediction was made.
Model-agnostic vs model-specific	Agnostic: works on any model by probing inputs and outputs (LIME, SHAP). Specific: uses a model's internals (gradients, tree paths).
Feature importance / attribution	A score for how much each input feature pushed the prediction up or down.
Surrogate model	A simple model (e.g. a small tree) trained to imitate a complex one, so the simple one can be read.
Saliency map / heatmap	An image overlay showing which pixels most influenced the decision.
Counterfactual / recourse	Smallest change that flips the result: "loan approved if income were ₹5,000 higher". Recourse = a change the person can actually make.
Faithfulness vs plausibility	Faithful (high fidelity): truly reflects how the model decided. Plausible: merely looks convincing. A plausible explanation can still be wrong.
Concept-based explanation	Explains using human concepts ("stripes", "tumour edge") instead of raw features or pixels.
Human-in-the-loop (HITL)	A human reviews or can override AI decisions, using explanations to judge them.
Responsible / Trustworthy AI	Explainability + fairness + privacy + robustness + accountability + safety.
Right to explanation	A person's legal right to a meaningful explanation of an automated decision about them (GDPR, EU AI Act Art. 86).

3. Main XAI Methods

Method	Type	Idea in one line	Typical use

Method	Type	Idea in one line	Typical use
Linear models, decision trees, rule lists	Intrinsic, global	Coefficients or if–then rules are the explanation.	Credit, clinical scores
GAMs / Explainable Boosting Machine	Intrinsic, global	Sum of one plottable curve per feature.	Tabular healthcare, finance
Permutation importance; PDP, ICE, ALE	Agnostic, global	Shuffle a feature and see accuracy drop; plot effect of one feature.	Any model
LIME; Anchors	Agnostic, local	Fit a small linear model (or if–then rule) around one example.	Text, tabular, images
SHAP (Kernel, Tree, Deep)	Agnostic + specific	Shapley values (game theory) share the prediction among features.	Industry standard
Saliency, Integrated Gradients, Grad-CAM, LRP	Gradient-based, local	Gradients/relevance flowing back highlight key pixels or words.	Medical imaging, vision
Counterfactuals (DiCE)	Agnostic, local	Search for the smallest change that flips the decision.	Loans, hiring
TCAV, Concept Bottleneck Models	Concept-based	Measure or enforce human concepts inside the model.	Vision, medicine
Prototypes (ProtoPNet), influence functions	Example-based	"This looks like that": explain via similar training cases.	Images, data debugging
Attention visualisation	Model-specific	Shows attended tokens; not always a faithful explanation.	Transformers, NLP

Worked example: SHAP in one line

SHAP writes every prediction as a base value plus one contribution per feature:

$$f(x) = \phi_0 + \phi_1 + \phi_2 + \cdots + \phi_M$$

Here ϕ_0 is the average prediction over all applicants and each ϕ_i is the push given by feature i. **Loan example:** average approval probability $\phi_0 = 0.30$; income +0.25; good credit history +0.15; existing debt −0.10. Prediction = 0.30 + 0.25 + 0.15 − 0.10 = **0.60**. The explanation: "income helped most; debt pulled the score down."

4. Explaining Large Language Models and Generative AI

Classic methods struggle with models of billions of parameters, so new research areas and buzzwords have emerged:

Buzzword	Meaning
Mechanistic interpretability	Reverse-engineering a network's internal computation into human-understandable parts, like a microscope for AI. An MIT Technology Review Breakthrough Technology of 2026.
Neurons, features, superposition	Models store more concepts (features) than neurons, so concepts overlap (superposition). Polysemantic neurons mix many concepts; monosemantic features mean one.
Sparse autoencoders (SAEs), dictionary learning, transcoders	Split tangled activations into millions of single-meaning features (e.g. a "Golden Gate Bridge" feature in Claude 3 Sonnet, 2024).
Circuits, attribution graphs, circuit tracing	Trace the chain of features from prompt to answer, revealing multi-step reasoning and planning inside the model.
Probing, logit lens, activation patching (causal tracing)	Probing: a small classifier tests if a layer stores some information. Logit lens: read each layer's guess. Patching: swap activations to find what causes a behaviour.
Steering vectors / representation engineering	Add a direction to internal activations to switch a concept on or off (e.g. less toxic).
Chain-of-thought (CoT) faithfulness and monitorability	Reasoning models write step-by-step "thoughts" that aid monitoring, but these do not always reveal the real reason.
Source attribution / grounding in RAG	RAG answers cite source documents, so users can verify claims and spot hallucinations.
Model cards, system cards, datasheets	Documents describing a model's intended use, data, evaluation and limits.

5. How Good Is an Explanation? (Evaluation)

Faithfulness: remove the features called important; if the prediction barely changes, the explanation was unfaithful (deletion/insertion curves). **Sanity checks:** randomise the weights; a trustworthy saliency map must change. **Stability:** small input noise should not change the explanation. **Human-centred:** do explanations help users decide better? Levels: application-grounded (experts, real task), human-grounded (lay users), functionally-grounded (proxy metrics).

6. Key Applications

Domain	How XAI is used
Healthcare	Grad-CAM heatmaps on X-ray, CT and retinal scans; SHAP for sepsis or readmission risk; checks that the model looks at the lesion, not a hospital tag.
Finance and insurance	Reason codes for credit refusals, fraud-alert explanations, model-risk validation, premium justification.
HR, law, public sector	Auditing hiring and risk-assessment tools for hidden bias (gender, caste or region proxies).
Autonomous systems, cybersecurity	Why a car braked; why traffic or a file was flagged, for fast analyst triage.
Industry, agriculture, climate	Predictive maintenance sensors; crop-disease leaf heatmaps; yield and weather drivers.
LLM safety, science, education	Detecting deception, jailbreaks or hidden goals in LLMs and agents; discovering biomarkers; explaining adaptive-learning recommendations.

7. Regulation and Standards

- **EU AI Act:** high-risk systems must be interpretable by users (Art. 13), allow human oversight (Art. 14) and give affected persons an explanation of individual decisions (Art. 86). Transparency duties for chatbots and AI-generated content (Art. 50) apply from 2 Aug 2026; after the 2026 Digital Omnibus, high-risk duties apply from 2 Dec 2027 (Annex III) and 2 Aug 2028 (Annex I).

- **GDPR:** Art. 22 / Recital 71 on automated decisions, the origin of the "right to explanation". **USA:** NIST AI RMF (2023); NIST Four Principles of XAI (explanation, meaningful, explanation accuracy, knowledge limits).
- **India:** MeitY India AI Governance Guidelines (Nov 2025); RBI FREE-AI framework for finance (2025). **Standard:** ISO/IEC 42001 AI management systems.

8. Popular Tools and Libraries

Purpose	Tools
General (tabular, text, image)	SHAP, LIME, ELI5, InterpretML (EBM), Alibi Explain, IBM AI Explainability 360 (AIX360), OmniXAI, Microsoft DiCE (counterfactuals)
Deep learning	Captum (PyTorch), tf-explain, pytorch-grad-cam, Zennit / iNNvestigate (LRP), Quantus (evaluating explanations)
LLMs and transformers	TransformerLens, nnsight, SAELens, Gemma Scope / Gemma Scope 2 (Google DeepMind), Neuronpedia, circuit-tracer (Anthropic, open-source), Inseq, BertViz, LIT (Learning Interpretability Tool)
Cloud platforms	Google Vertex AI Explainable AI, Amazon SageMaker Clarify, Azure Machine Learning Responsible AI dashboard, IBM watsonx.governance
Fairness (companion to XAI)	Fairlearn, IBM AI Fairness 360, Google What-If Tool

9. Challenges and Pitfalls

- **Accuracy–interpretability trade-off:** often overstated; Rudin (2019) argues interpretable models are frequently as accurate for high-stakes tabular tasks.
- **Unfaithful or manipulated explanations:** LIME/SHAP can be gamed to hide bias ("fairwashing"); attention and chain-of-thought may mislead.
- **Rashomon effect and explainer disagreement; correlation is not causation;** scale and cost for billion-parameter models; over-trust (automation bias).

10. Landmark Research to Date

Year	Work	Why it matters
2016	LIME: Ribeiro, Singh & Guestrin, KDD	Popularised local, model-agnostic explanations.
2017	SHAP (Lundberg & Lee, NeurIPS); Integrated Gradients (ICML); Grad-CAM (ICCV); DARPA XAI; Doshi-Velez & Kim	Shapley-value standard; gradient heatmaps; XAI becomes a funded, evaluated field.
2017–18	Counterfactual explanations: Wachter, Mittelstadt & Russell	Linked explanations to GDPR and actionable recourse.
2018	TCAV (Kim et al., ICML); Sanity Checks for Saliency Maps (Adebayo et al., NeurIPS)	Concept-level explanations; showed some saliency maps ignore the model.
2019	Rudin, "Stop explaining black box models…", Nature Machine Intelligence	Case for interpretable-by-design models in high-stakes use.
2020	Concept Bottleneck Models (Koh et al., ICML); "Fooling LIME and SHAP" (Slack et al., AIES)	Concept-based design; exposed manipulation risk of post-hoc methods.
2022	Toy Models of Superposition (Anthropic); ROME: locating and editing facts in GPT (Meng et al., NeurIPS)	Explained why neurons are polysemantic; found where facts are stored.
2023–24	Towards / Scaling Monosemanticity (Anthropic); Gemma Scope (Google DeepMind)	Sparse autoencoders extract millions of interpretable features from production LLMs.
2025	Circuit tracing, "On the Biology of a LLM" and open-source circuit-tracer (Anthropic); "Reasoning Models Don't Always Say What They Think" (Chen et al.); Gemma Scope 2	Traced reasoning inside LLMs; showed chain-of-thought can be unfaithful; open tools for Gemma 3.
2026	Mechanistic interpretability: MIT Tech Review Breakthrough Technology; EU AI Act Art. 50 in force	Interpretability moves from research to safety practice and law.

11. Key Takeaways

(1) Prefer an interpretable model when it is accurate enough. **(2)** Otherwise use post-hoc methods (SHAP, LIME, Grad-CAM, counterfactuals), and check them for faithfulness. **(3)** For LLMs, mechanistic interpretability (SAEs, circuit tracing) is the fastest-growing frontier. **(4)** Explanations must suit the audience: engineers, doctors, customers and regulators need different ones. **(5)** Explainability is now a legal and ethical requirement, not an optional extra.