Why we cannot ask a model why

A history of explainable AI, from decision trees to deep learning, and why the legal right to an explanation arrived just as systems stopped being able to give one.

> Explainability is not a property of a model. It is a relationship between a model and the person who has to live with its decision.

HOW IT GOT HERE

1998 — Interpretable by construction

Decision trees and rule lists dominate applied machine learning. A prediction can be read off the path that produced it.

2012 — The accuracy trade

Deep networks take the accuracy lead on image recognition. The representations that win are distributed across millions of weights and legible to nobody.

2016 — Post-hoc explanation

LIME, and SHAP the year after, propose explaining a prediction by approximating the model locally rather than opening it.

2017 — DARPA XAI

A dedicated research programme funds work on models that explain themselves, on the premise that opacity is a defence liability.

2018 — A right to an explanation

GDPR takes effect. Article 22 restricts solely automated decisions with legal or similarly significant effects, and requires meaningful information about the logic involved.

2019 — Federal footing

The AI in Government Act directs federal attention to how agencies adopt and account for AI systems.

2022 — Principles without teeth

The Blueprint for an AI Bill of Rights states that people should get notice and a plain-language explanation. It binds nobody.

2026 — Disclosure by statute

State frontier-AI law arrives with published safety protocols and incident disclosure: explanation moves from a research problem to a filing requirement.

THE TRADE THAT CREATED THE PROBLEM

For most of the history of applied machine learning, explanation was free. A decision tree is a sequence of questions, and the answer to a prediction is the path taken through it. A linear model is a list of weights: the reason for an output is the coefficients that produced it. Nobody wrote papers on interpretability because interpretability was not separable from the model.

That ended when the accuracy of distributed representations overtook the accuracy of legible ones. A deep network does not store a rule about what a tumour looks like. It stores a very large number of weights whose joint behaviour approximates one, and there is no location in the model where the rule can be read off. The field did not choose opacity. It chose accuracy, and opacity was attached to it.

Everything since has been an attempt to buy back what that trade gave away, and the two available routes are unattractive in different ways. Either build models that are interpretable by construction and accept the accuracy cost, or build the opaque model and explain it after the fact, accepting that the explanation is an approximation of the model rather than the model itself.

WHAT POST-HOC EXPLANATION ACTUALLY GIVES YOU

The dominant techniques - locally interpretable approximations, and Shapley-value attributions borrowed from cooperative game theory - answer a narrow question well: which input features, for this particular prediction, moved the output most. That is genuinely useful. It catches the model that classifies skin lesions by the presence of a surgical ruler in the frame, because the ruler appears next to lesions a clinician already suspected. It catches the credit model whose strongest feature is a postcode.

It does not tell you why the model believes what it believes, because the model does not believe anything. An attribution is a statement about sensitivity: change this input, and the output moves this much. Sensitivity is not reasoning, and treating it as reasoning is where explanation becomes a liability rather than a safeguard.

The practical failure mode is an explanation that satisfies the person receiving it without being true of the system. A plausible-sounding attribution is very hard to contest, which is precisely the property you do not want in a document that exists to let someone contest a decision.

THE RIGHT TO AN EXPLANATION ARRIVED AT THE WRONG TIME

European data protection law established that a person subject to a solely automated decision with legal or similarly significant effects is entitled to meaningful information about the logic involved. The obligation is sound. It landed in the same decade that the systems making those decisions became the least explicable they had ever been.

That mismatch has been absorbed rather than resolved. In practice the obligation is met with a description of the categories of data used and the general shape of the process, which is compliance rather than explanation. The person still cannot tell why they specifically were refused.

American federal attempts have taken a different route, requiring assessment rather than explanation: document how the system performs and on whom, and report it. That is a lower bar in one sense and a higher one in another. It does not owe the individual an account of their own decision, but it does force a claim about the system's behaviour that can be checked.

LONGTERMISM AND THE ARGUMENT ABOUT WHICH HARM COUNTS

Two constituencies argue for explainability and mean different things by it. One is concerned with catastrophic risk from systems substantially more capable than current ones, and treats interpretability as a safety precondition: you should not deploy what you cannot inspect. The other is concerned with harm that is already occurring - denied claims, wrongful arrests from face recognition, credit and housing decisions that reproduce discrimination - and treats explanation as due process.

The disagreement is not really about the technique. It is about which harms are urgent, and therefore where finite research and regulatory attention should go. Framed as a choice it is a bad one, because the mechanism that lets a regulator audit a deployed benefits system is largely the mechanism that would let anyone audit a much more capable one.

Our position is that the near-term case is the stronger place to start, for a reason that is procedural rather than moral: harms already occurring generate records, complainants and evidence. A regime built to handle those has something to test itself against. A regime built only for anticipated harm has nothing.

WHAT WE THINK IS WORTH REQUIRING

Notice that an automated system was involved, at the point of the decision rather than in a policy document. A person cannot contest what they do not know happened.

A route to a human with the authority to change the outcome. Review that cannot reverse anything is not review.

Documented assessment of the system's performance, disaggregated across the groups it is used on, retained and disclosable. This is the only requirement in the list that produces evidence.

Restraint about explanation itself. Requiring an explanation for every automated decision produces a great many explanations of uncertain truth. Requiring notice, review and assessment produces fewer documents and more accountability.

SOURCES

01 Blueprint for an AI Bill of Rights - White House Office of Science and Technology Policy

02 General Data Protection Regulation, Article 22 - European Union

03 Explainable Artificial Intelligence (XAI) programme - DARPA

04 Algorithmic Accountability Act of 2023 - U.S. Congress

05 AI in Government Act of 2019 - U.S. Congress