THE BLACK BOX PROBLEM

What It Is, Why It Persists, and How to Engineer Around It

Opacity in machine learning is not one problem but three. Only one of them resists a fix by disclosure.

Why the model resists explanation, what it costs, and the architecture that doesn't wait for it to change.

01 Three forms of opacity, only one is intrinsic

02 The technical reasons models resist explanation

03 What it costs when a model fails silently

04 What the law now requires you to evidence

05 The architecture that doesn't depend on interpretability

26,000

Families flagged by an opaque fraud-risk model on grounds they could not see or contest

The Dutch tax authority's childcare benefits system used nationality as a risk indicator. Over a thousand children were removed from their homes. Source: Dutch Data Protection Authority.

$304M

Zillow Offers inventory write-down after its valuation model failed under changed market conditions

Q3 2021, followed by closure of the business line and a headcount reduction of around 25 per cent.

~25%

Share of prompts for which Anthropic's own interpretability tooling produced a full explanation

Circuit tracing on Claude 3.5 Haiku is the most advanced published mechanistic interpretability work to date. That is the state of the art, not a gap in effort.

Three forms of opacity. Only one is intrinsic to the technology.

Opacity by intent

Architecture, weights, or training data withheld as confidential. A contractual problem, soluble by negotiation or disclosure.

Opacity by expertise gap

The information exists but the organisation lacks the capability to read it. A resourcing problem, soluble by hiring.

Opacity by construction

No party holds the explanation, including the developer. Not soluble by disclosure. This is the substantive problem.

Why opacity is structural, not an oversight.

Distributed representation

Concepts are directions in activation space, not neurons. A single unit participates in many unrelated computations.

No specification to verify against

A model has a training objective, not a requirement decomposition. There is only a distribution to test, not a document to check.

Unfaithful self-explanation

A model's stated reasoning is generated by the same process as its answer. It is not a record of what actually determined the output.

An opaque model is not disqualified from a system. It is disqualified from being the only thing standing between a decision and its consequence.

THE BLACK BOX PROBLEM, SECTION 8

Five moves that don't wait for interpretability to mature.

- ✓ **Keep opaque components off the actuation path**
  Place a deterministic, verifiable envelope between the model and any physical or entitlement action.
- ✓ **Instrument the input distribution independently**
  Detect drift from outside the model. A failing model does not report that it has begun to fail.
- ✓ **Require counterfactual explanation for personal decisions**
  States the minimal input change that would flip the outcome, without disclosing model internals.
- ✓ **Log the decision, not just the model**
  A model card describes the artefact. Only decision-level lineage lets you reconstruct a specific past outcome.
- ✓ **Scale transparency to consequence, not complexity**
  A routing classifier and a benefit determination do not warrant the same evidentiary standard.

This is a systems architecture problem, not a research problem.

Read the full piece for the technical origins, the legal exposure under UK and EU law, and the mitigation architecture that doesn't depend on full interpretability.

Terence Kok
AI & Innovation Strategist

SAVE · COMMENT · FOLLOW