Category

Data & AI

AI Transparency, Explainability & Model Interpretation Playbook

Choosing explainability appropriate to the decision's stakes, not the maximum available accuracy

- Practitioner
- Advanced
- Template Included

- Administrator
- 03 September 2019
- 3 minute(s)

Overview

A framework for AI transparency, explainability, and model interpretation — selecting explainability approaches proportional to decision stakes, distinguishing genuine explanation from post-hoc rationalization, and building interpretation capability appropriate to different stakeholder audiences from technical teams to affected individuals.

Should organizations always choose the most explainable AI

approach available, even at some cost to accuracy? Not universally — the right explainability-accuracy trade-off depends on decision stakes; a low-stakes internal recommendation might reasonably prioritize accuracy, while a consequential decision affecting individuals

(credit, employment) typically warrants prioritizing genuine explainability even at some accuracy cost.

What's the difference between genuine explanation and post-hoc

rationalization in AI interpretability? Genuine explanation reflects what actually drove the model's decision; post-hoc rationalization techniques can produce plausible-sounding explanations that don't accurately reflect the model's actual internal decision process — a distinction that matters significantly for high-stakes explainability requirements.

{ "@context": "https://schema.org", "@type": "FAQPage", "mainEntity": [{ "@type": "Question", "name": "Should organizations always choose the most explainable AI", "acceptedAnswer": { "@type": "Answer", "text": "approach available, even at some cost to accuracy?\nNot universally — the right explainability-accuracy trade-off depends\non decision stakes; a low-stakes internal recommendation might\nreasonably prioritize accuracy, while a consequential decision affecting\nindividuals (credit, employment) typically warrants prioritizing\ngenuine explainability even at some accuracy cost." } }, { "@type": "Question", "name": "What's the difference between genuine explanation and post-hoc", "acceptedAnswer": { "@type": "Answer", "text": "rationalization in AI interpretability?\nGenuine explanation reflects what actually drove the model's decision;\npost-hoc rationalization techniques can produce plausible-sounding\nexplanations that don't accurately reflect the model's actual internal\ndecision process — a distinction that matters significantly for\nhigh-stakes explainability requirements." } }] }

Subscriber access

Unlock this playbook

This playbook — including every framework, template, and step-by-step section — is available free to Think Insights subscribers. Enter your email to unlock it instantly and get our weekly insights newsletter. No account needed, and access is remembered on this device.

First Name

Last Name

Email

Country - Select -

■ I agree to data processing in accordance with GDPR

Submit

Leave this field blank

References

Related playbooks

Data & AI

AI in Data Engineering & Pipelines Orchestration Playbook

A framework for using AI in data engineering and pipeline orchestration — AI-assisted pipeline code generation, automated data quality monitoring, and intellige

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

AI Data Readiness & Foundation Models Input Quality Playbook

A framework for assessing and building AI data readiness — data quality, completeness, and structure appropriate for foundation model input — recognizing that A

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

AI for Design, Prototyping & Experimentation Workflows Playbook

A framework for using AI in design, prototyping, and experimentation workflows — accelerating prototype generation and iteration speed while maintaining genuine

- Practitioner
- Intermediate
- Template Included

Read playbook

Talk to us about AI transparency and explainability

Tags

- AI Transparency
- Explainability
- Model Interpretation

Author

Administrator

Think Insights Administrator