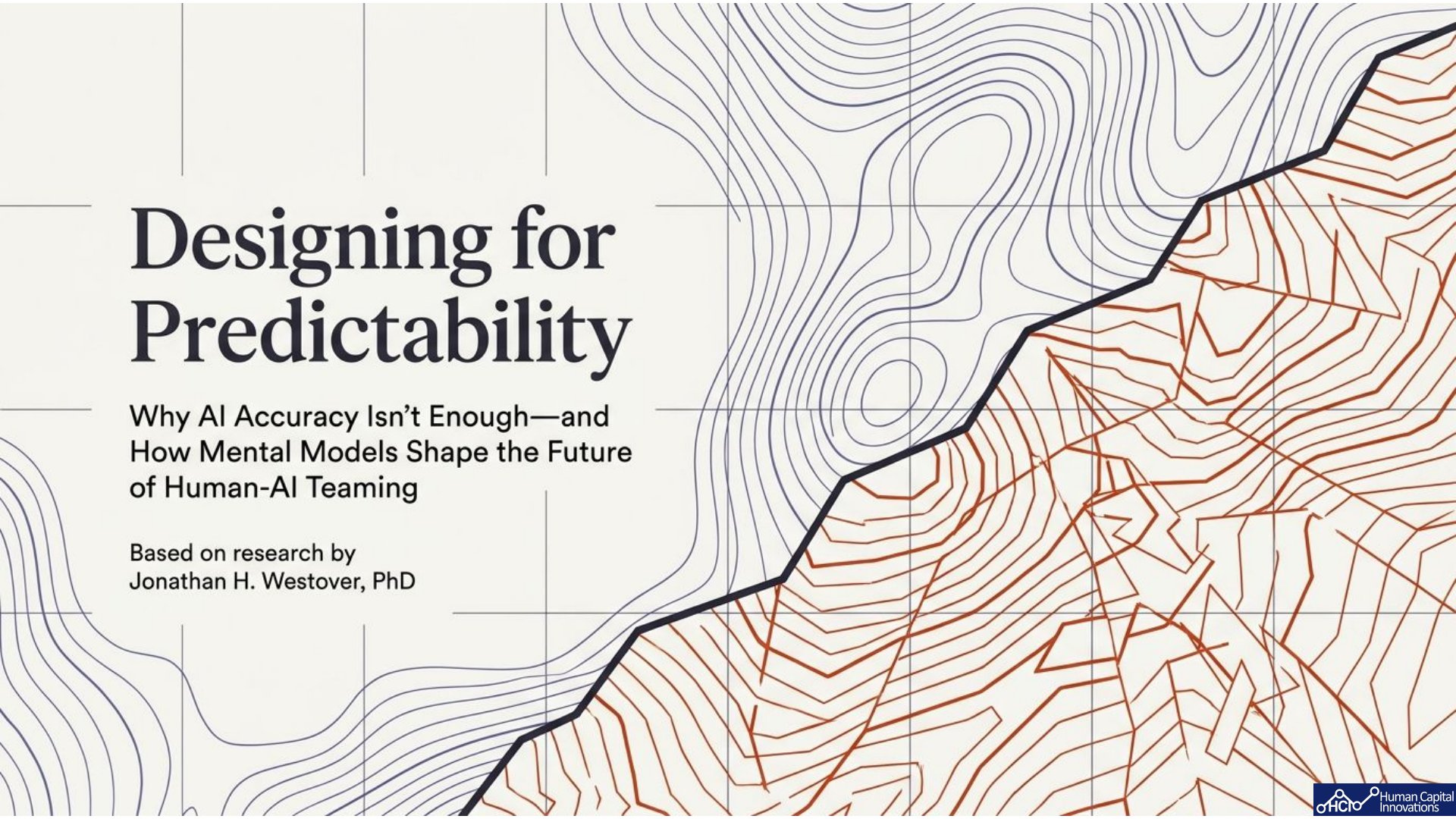




Designing for Predictability

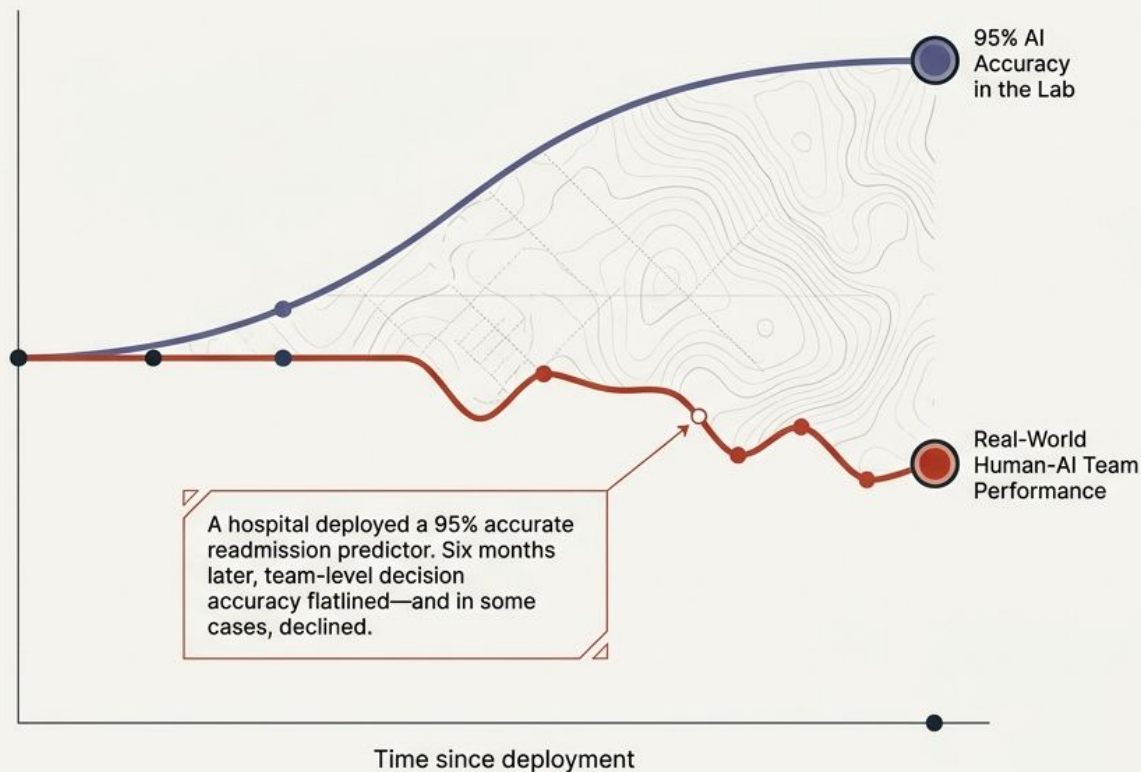
Why AI Accuracy Isn't Enough—and
How Mental Models Shape the Future
of Human-AI Teaming

Based on research by
Jonathan H. Westover, PhD

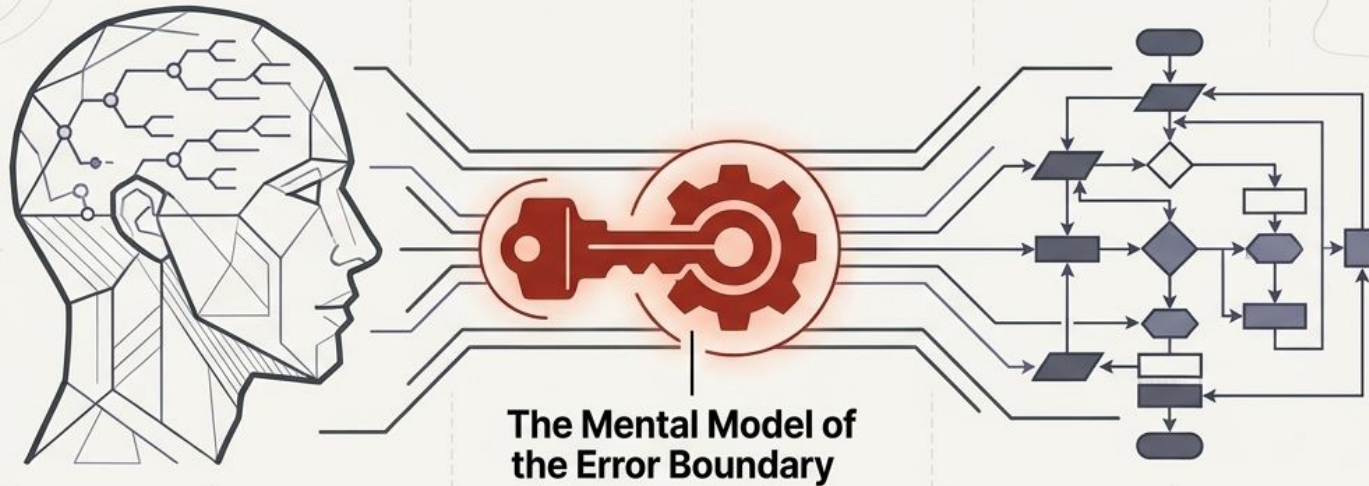
If we build a more accurate model, decision outcomes will improve. It sounds reasonable. And it is dangerously incomplete.

The Dangerous Trap of High-Lab Accuracy

The Accuracy Illusion: Diverging Trajectories of Lab vs. Real-World Performance



The Mental Model of the Error Boundary: Mapping the Trust Zone

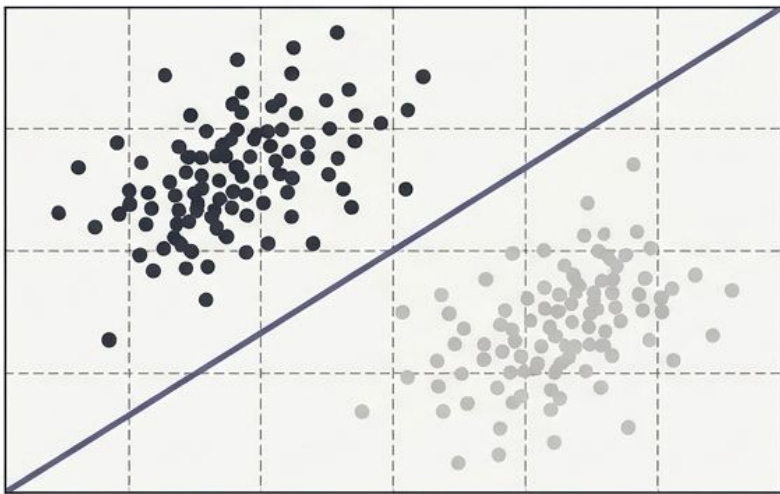


Meaningful oversight is impossible if humans cannot distinguish an AI's zones of competence from its zones of failure.

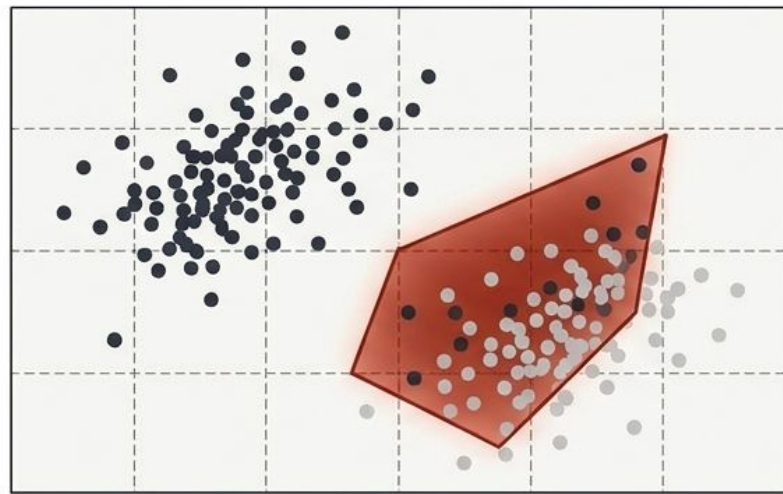
The only job that justifies keeping a human in the loop is knowing when to trust the AI and when to override it.

The Cartography of Trust: Mapping the AI Error Boundary

What the AI Predicts



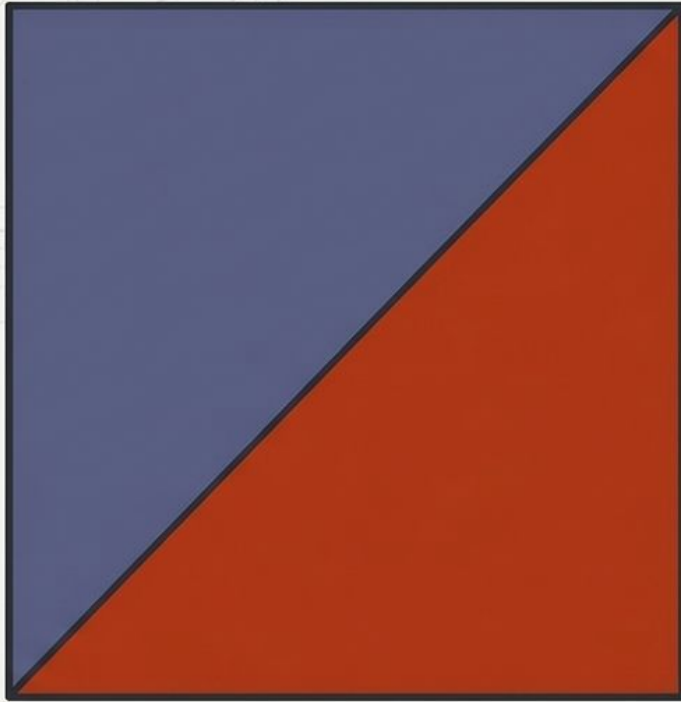
Where the AI Fails



Stop asking: What does it predict?
Start asking: For what kinds of inputs does this AI make mistakes?

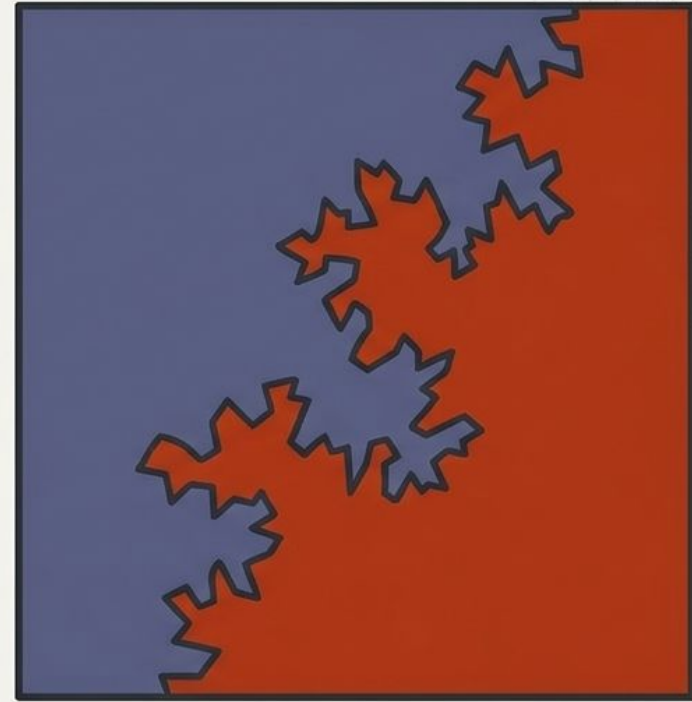
	AI is Correct	AI is Incorrect
Human Trusts		Misplaced Trust Human inherits the AI's error.
Human Overrides	Misplaced Distrust Human forfeits the AI's value.	

In domains like healthcare and criminal justice, the penalty for an incorrect action heavily outweighs the reward for a correct one. Unpredictable AI actively degrades professional confidence and harms vulnerable populations.



Learnable Boundary

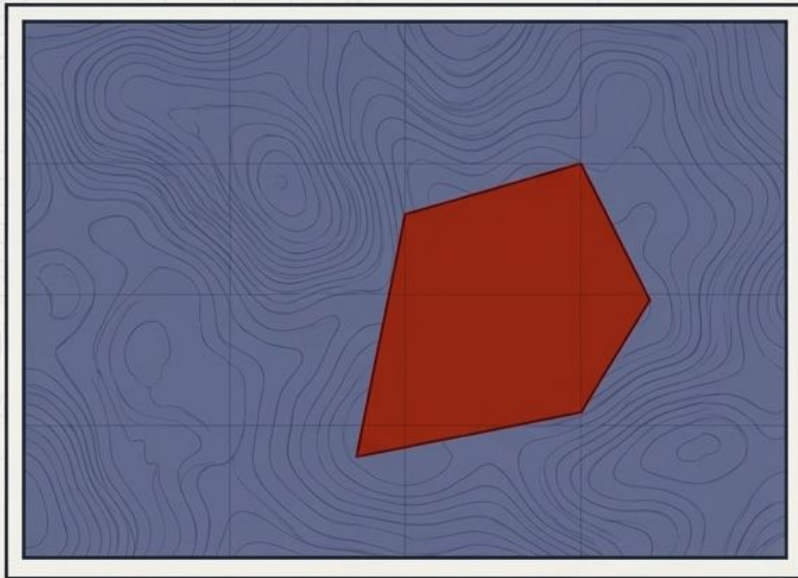
Pillar 1: Parsimony. The failure pattern must be describable with minimal rules.



Unlearnable Boundary

Actionable Rule: Accept a modest machine-learning accuracy trade-off if it drastically simplifies the error boundary for the human team.

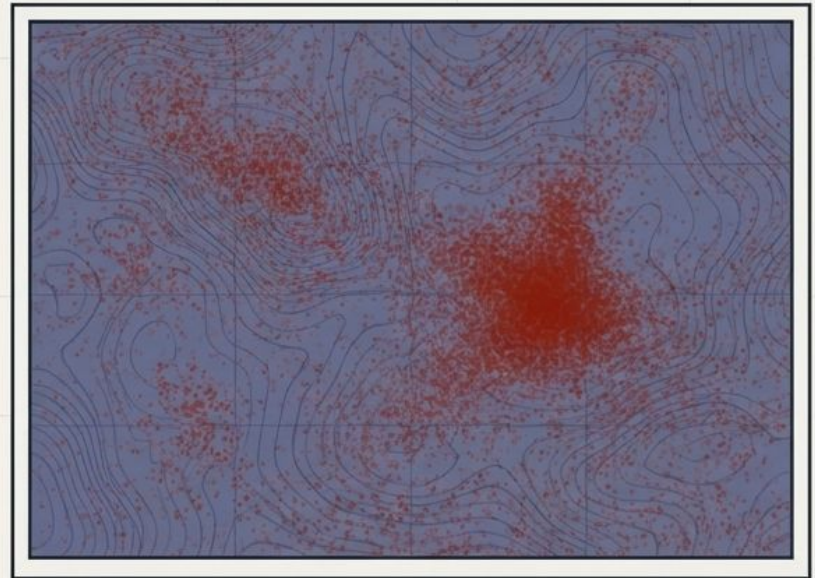
The Cartography of Trust: Pillar 2 - Non-Stochasticity



Non-Stochastic Error Boundary

Pillar 2: Non-Stochasticity

The AI must consistently fail for identifiable inputs. Stochastic (fuzzy) error boundaries destroy trust calibration because users cannot map a reliable pattern.

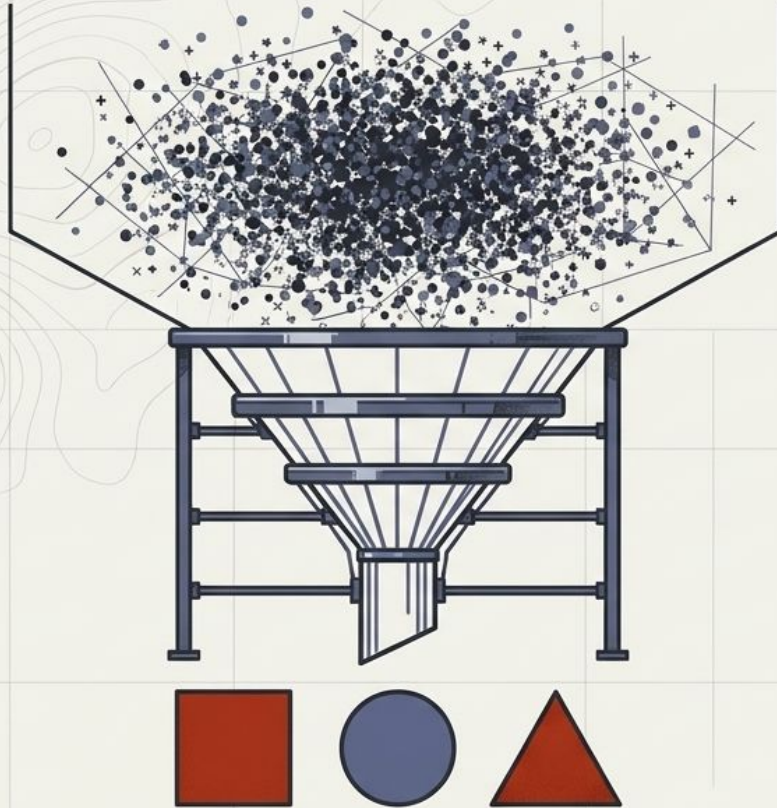


Stochastic Error Boundary

Actionable Rule: Prioritize confidence calibration

Communicate uncertainty transparently rather than presenting binary yes/no outputs. Set explicit expectations upfront.

Machine Scale: 1,000+ Features



Human Scale: 3 Clear Features

Pillar 3: Task Dimensionality

As human-visible features increase, the search space for discovering error patterns grows combinatorially.

Actionable Rule:

Align machine-visible and human-visible features.

Layer information presentation to prevent cognitive overload.

Adding features might improve solo AI accuracy while destroying team performance.

The Cartography of Trust: Comparative Case Study

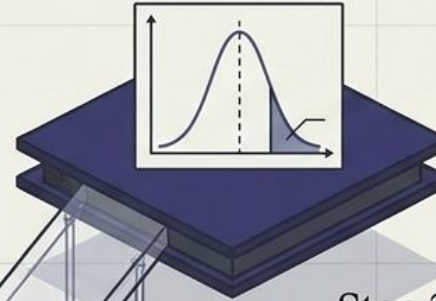
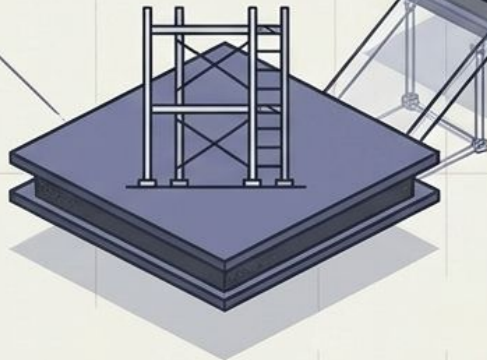
	Microsoft/Caruana Pneumonia GAM (The Winner)	COMPAS Recidivism System (The Failure)
Design	 Parsimonious, intelligible.	 High dimensionality, unlearnable, opaque.
Human-AI Team Outcome	High team performance. Clinicians easily mapped the error boundary, revealing a dangerous 'asthma anomaly' hidden in the data.	Actively harmed judges' trust calibration. Systematic errors were unmapped. Simple 2-feature models matched its predictive accuracy.

Pillar 3: Activation Staircase

The journey towards trust requires deliberate progression through increasing levels of exposure to AI capabilities and limitations.

Step 1: Scaffolded Learning.

Start with simple, clear cases where the error boundary is obvious, then progressively introduce complex scenarios.



Step 3: Representative Exposure.

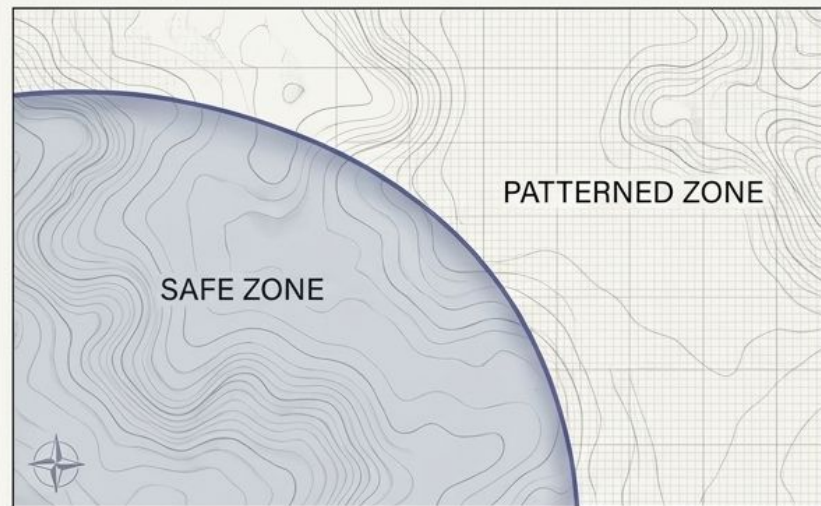
Expose users to the actual distribution of successes and failures they will encounter in the

Step 2: Explicit Failure Libraries.

Do not rely on trial and error. Curate documented examples of cases where the AI is known to fail, categorized by key features.

The Disruption of the Upgrade: Shifting Trust Boundaries

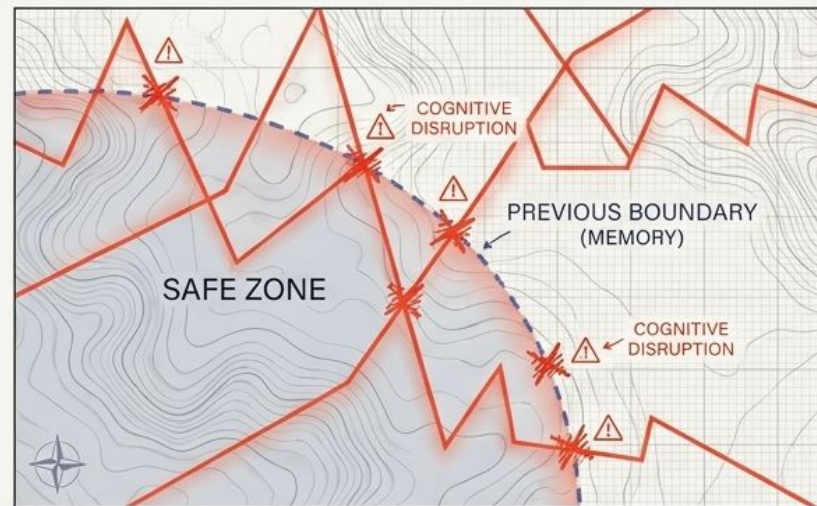
Before Update



The Danger of the 'Upgrade'

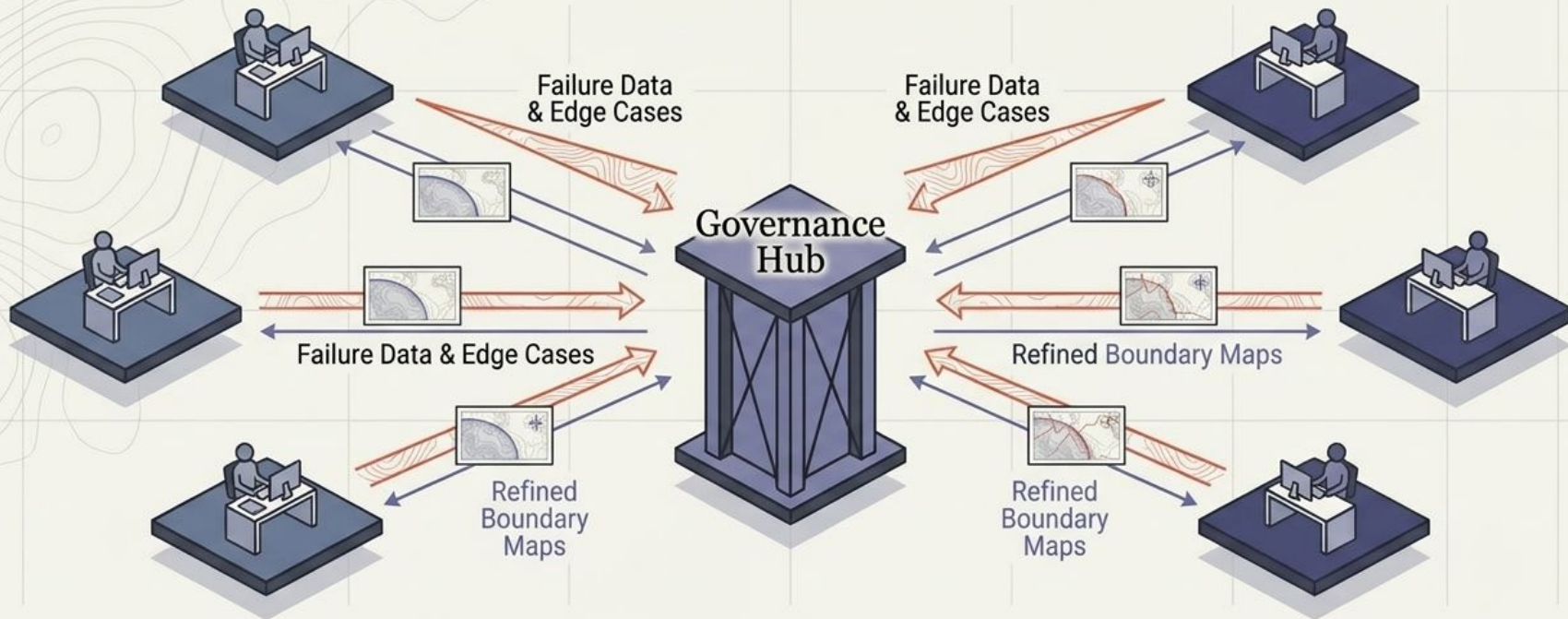
A more accurate model can actively harm team performance if it introduces new errors in unexpected places.

After Update



Trust Events

Model updates are not just code deployments; they are "trust events." Unmanaged, non-backward-compatible updates break established mental models and erode reliance.

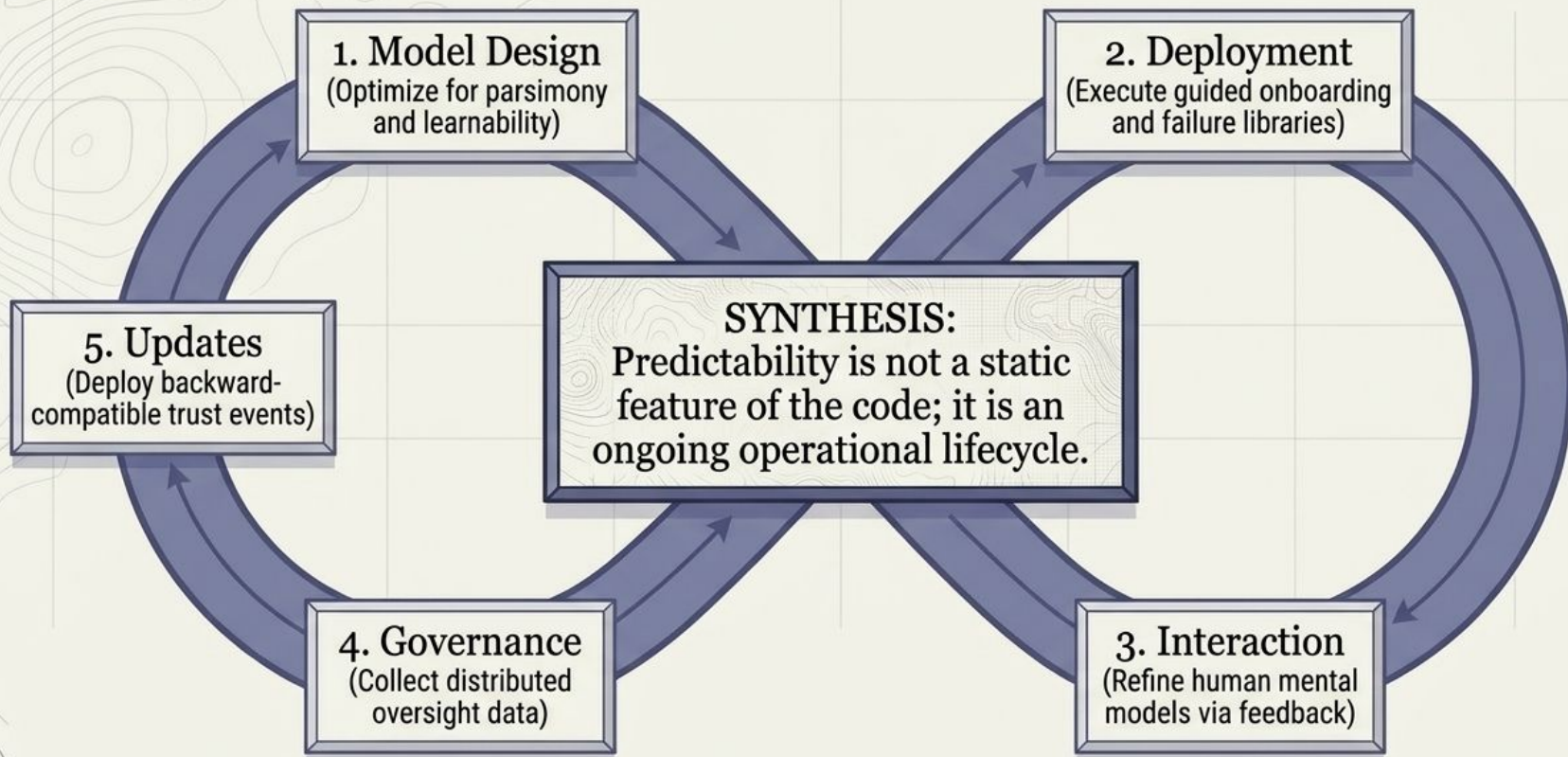


Distributed Oversight

No single user can map the entire error boundary across diverse populations and edge cases.

Collaborative Architecture

Establish 'scratchpad tools' and incident reporting systems to aggregate insights. Treat AI error boundary monitoring as a shared, continuous organizational capability.



Executive Playbook: The Cartography of Trust in Human-AI Systems

1



Optimize for human-AI team performance, not isolated AI accuracy.

2



Design learnable, parsimonious, and non-stochastic error patterns.

3



Invest in explicit failure-case training and guided onboarding.

4



Manage updates as trust events by prioritizing backward compatibility.

5



Build distributed governance systems to monitor boundaries continuously.

**The era of AI-advised decision making is here.
Success depends not just on building smarter
AI, but on building AI that humans can
understand well enough to use wisely.**