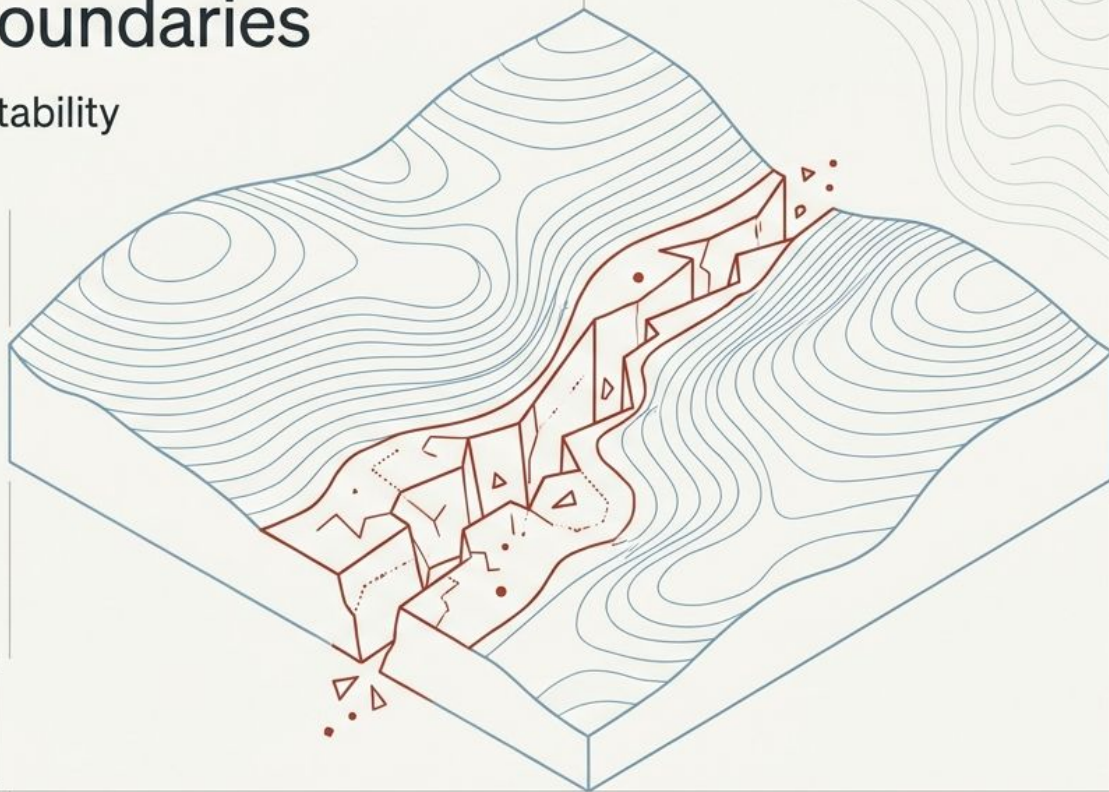


Engineering Human-AI Collaboration by Mapping Error Boundaries

Accuracy is necessary, but predictability unlocks team performance



Based on foundational research by Bansal et al. (2019) and insights from high-stakes AI deployments in healthcare and criminal justice.



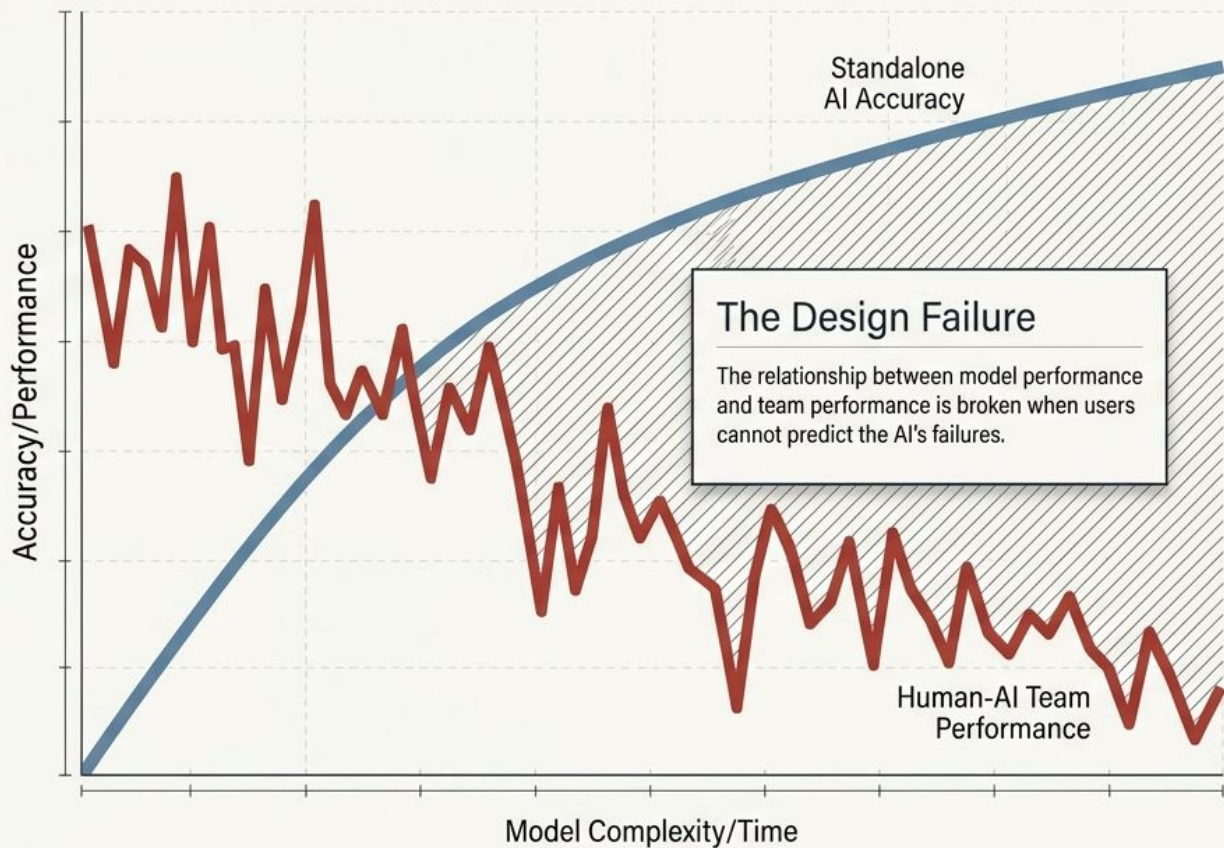
The Expectation:

Organizations assume AI accuracy directly drives value.



The Reality:

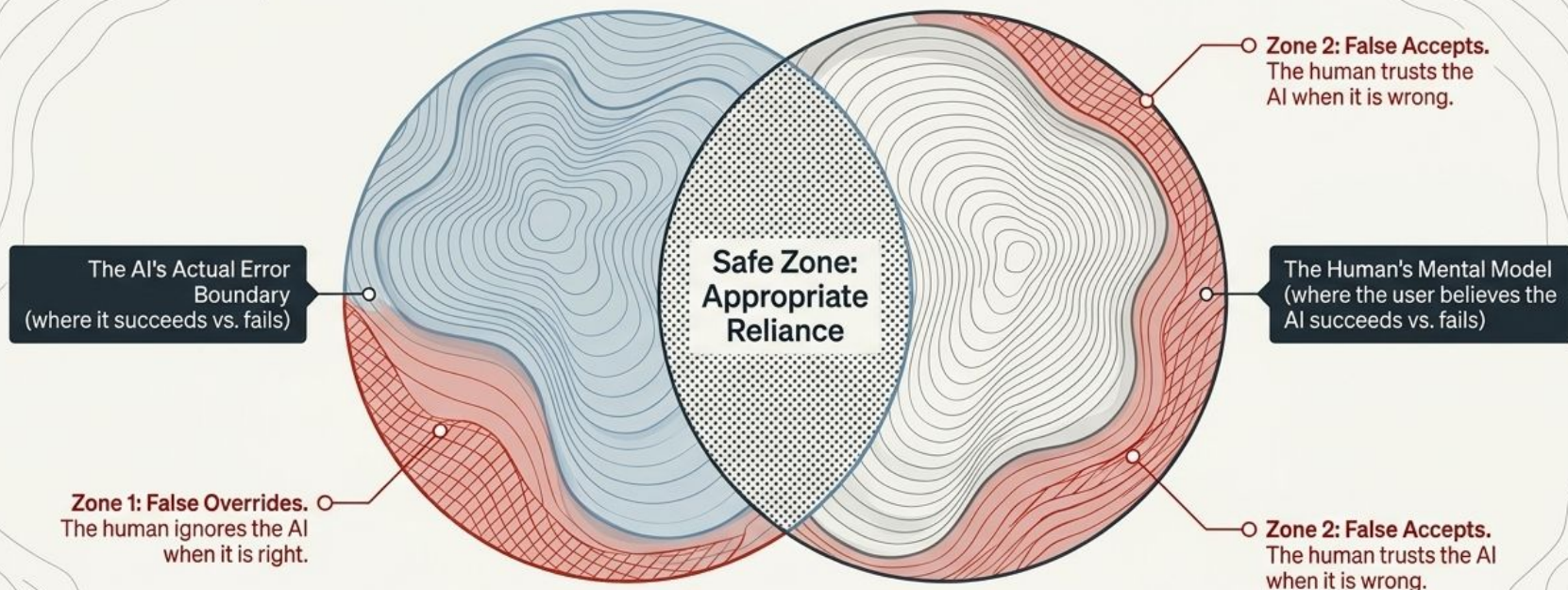
Skilled operators routinely ignore correct AI flags or accept incorrect ones when the system's logic is unpredictable.



Define the Error Boundary:

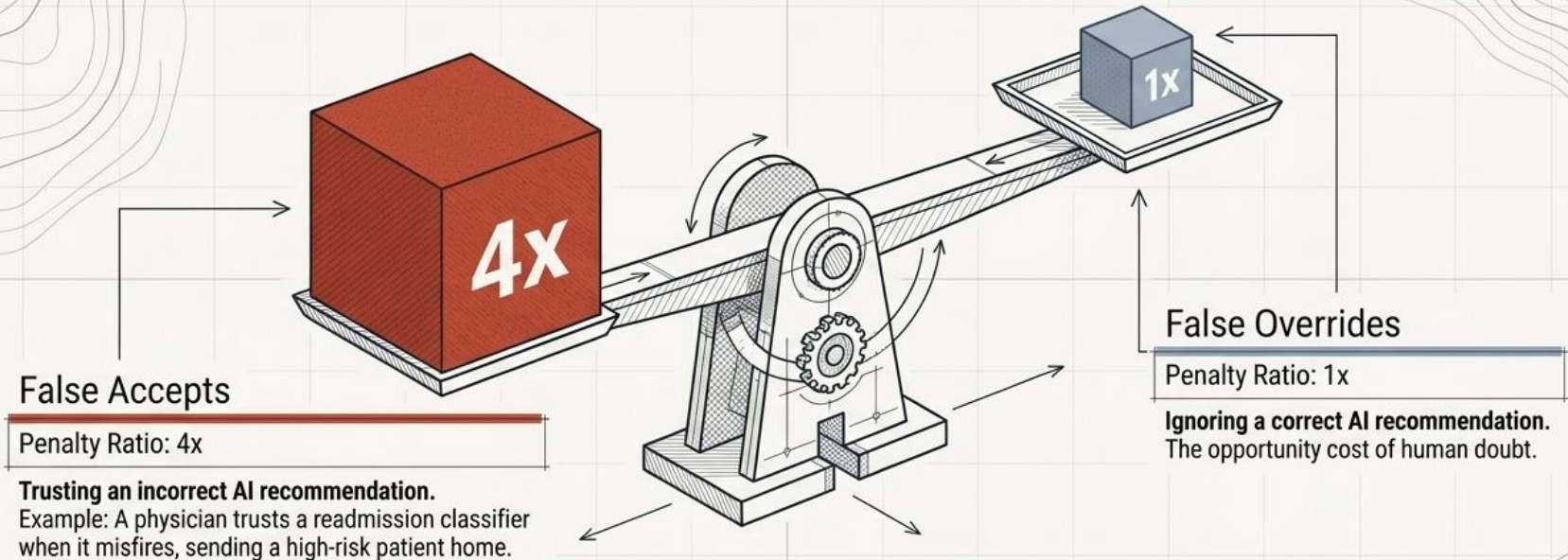
The mathematical pattern distinguishing where the system succeeds from where it fails.

The Missing Link is the Human's Mental Model of the Error Boundary



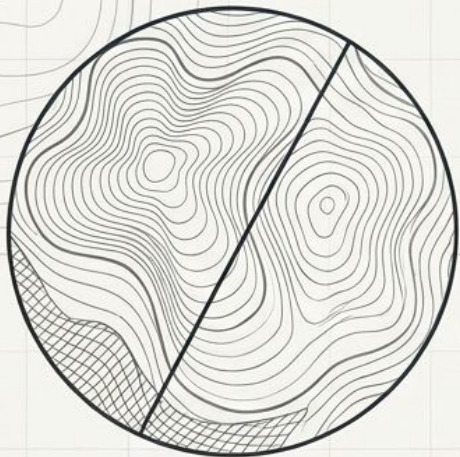
Core Insight: The goal is not global trust, but situational discernment.

The Asymmetric Cost of Misalignment in High-Stakes Environments



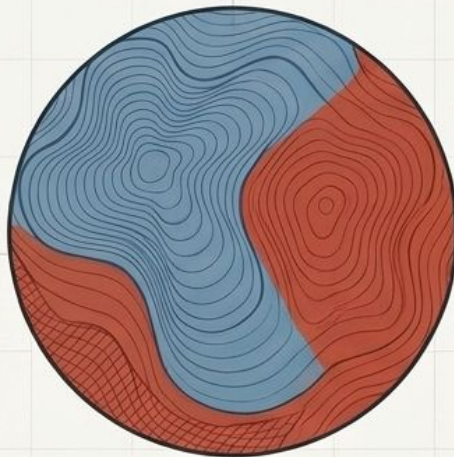
Bansal et al. (2019) demonstrated a 4:1 penalty ratio in controlled studies. In healthcare, finance, and criminal justice, **false accepts are exponentially more devastating than false overrides**. An unlearnable **error boundary** guarantees **false accepts**.

Three System Properties Dictate Error Boundary Learnability



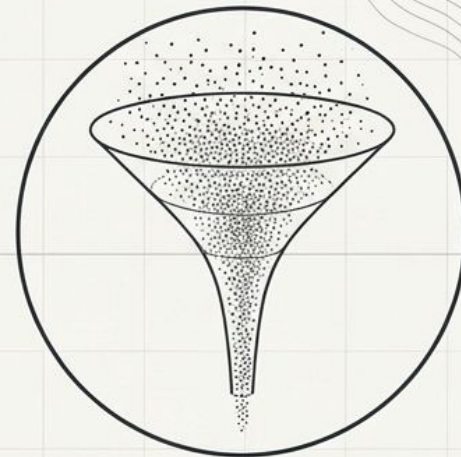
1. Parsimony

Simplicity of the error pattern.
Can the human describe the failure zone simply?



2. Non-Stochasticity

Predictability of the error.
Does the AI fail consistently, or randomly?



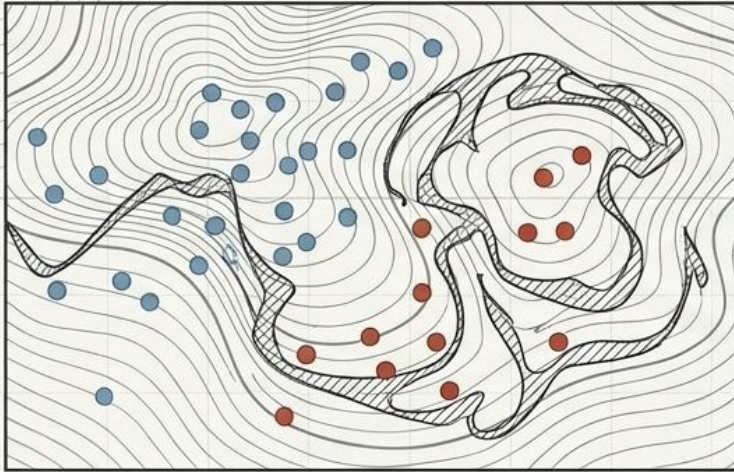
3. Dimensionality

Cognitive load management.
Can the human process the number of features required to spot the error?

Optimizing for these three pillars often requires **trading away fractional points of raw accuracy** to unlock massive gains in team performance.

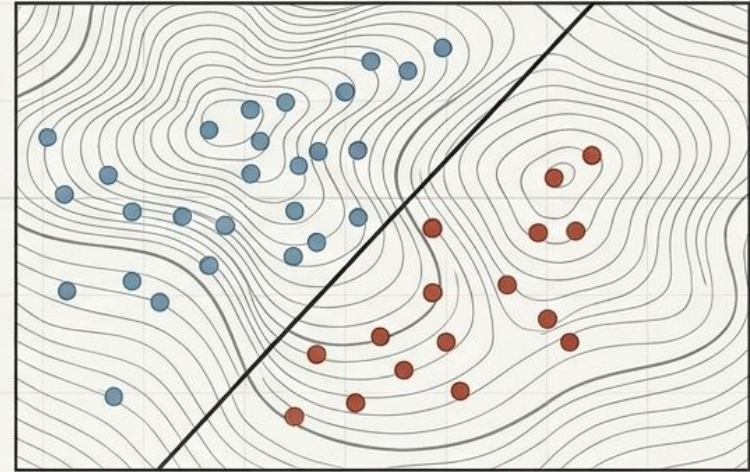
Pillar 1: Parsimonious Error Boundaries Outperform Complex Ones

High Accuracy, Low Parsimony



The Trap of Complexity: Bansal et al. found that multiple disjunctive conditions (e.g., fails on red and square OR small and circle) degrade team performance.

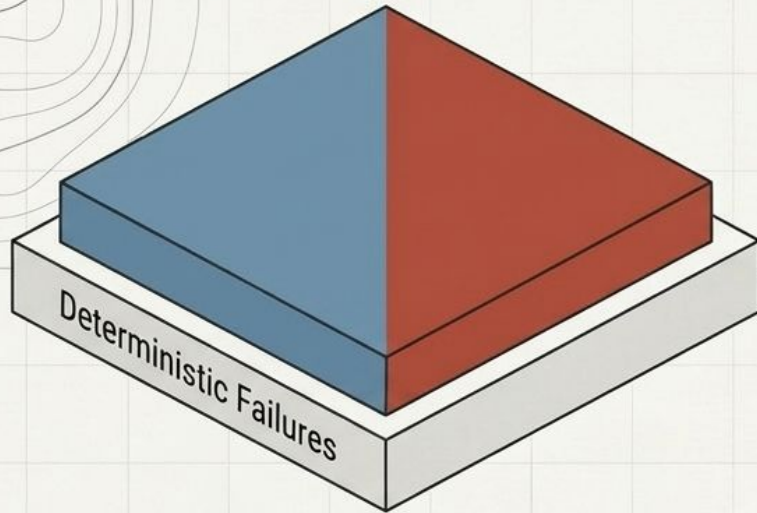
Lower Accuracy, High Parsimony



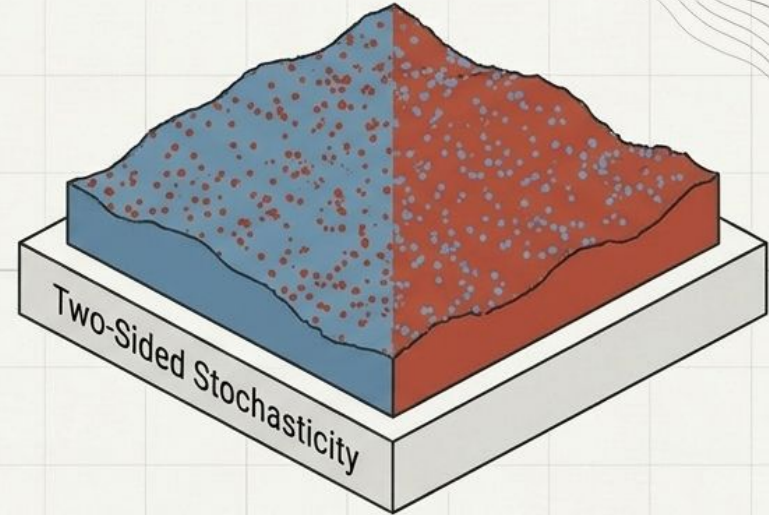
The Power of Simplicity: A single-conjunction boundary allows the human to build a reliable internal rule.

Diagnostic Question: Can you describe your model's failure zone in a single sentence?

Pillar 2: Stochasticity and Noise Destroy Mental Model Formation



Design Imperative: It is better for an AI to always fail in a specific region than to fail unpredictably across the board.



The Coin-Flipping Effect: When an AI succeeds in its weak zone and fails in its strong zone, human learning stalls completely.

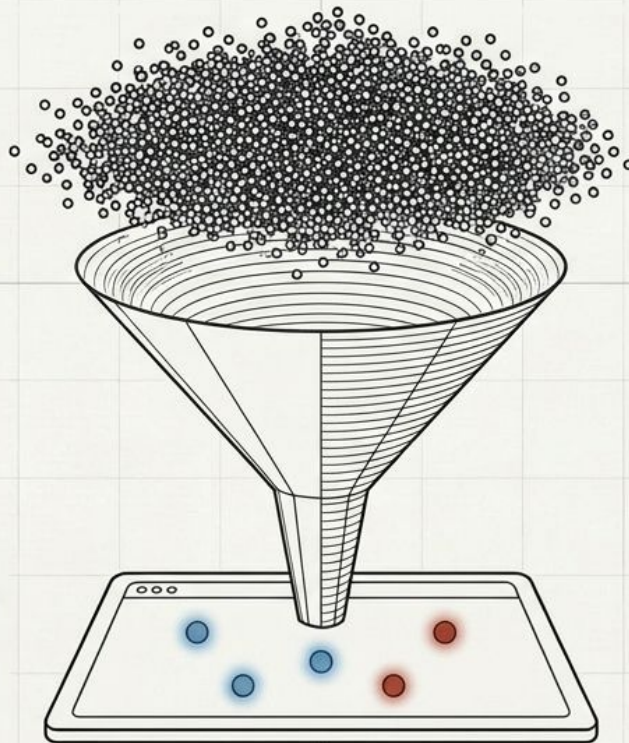


The Data: In experiments with 100 rounds of interaction, users facing stochastic errors showed near-zero improvement, reverting to 50% accuracy (guessing).

Pillar 3: High Task Dimensionality Breaks Human Pattern Recognition

The Paradox of Features: Adding features often improves standalone machine accuracy but obscures the error pattern from the human.

The Solution: Progressive disclosure. The AI can compute a thousand features internally, but the interface must curate only the most salient features necessary to spot the error boundary.



1,000+ Internal Model Features

Cognitive Overload: Human reasoning collapses under high-dimensional data (Poursabzi-Sangdeh et al., 2021).

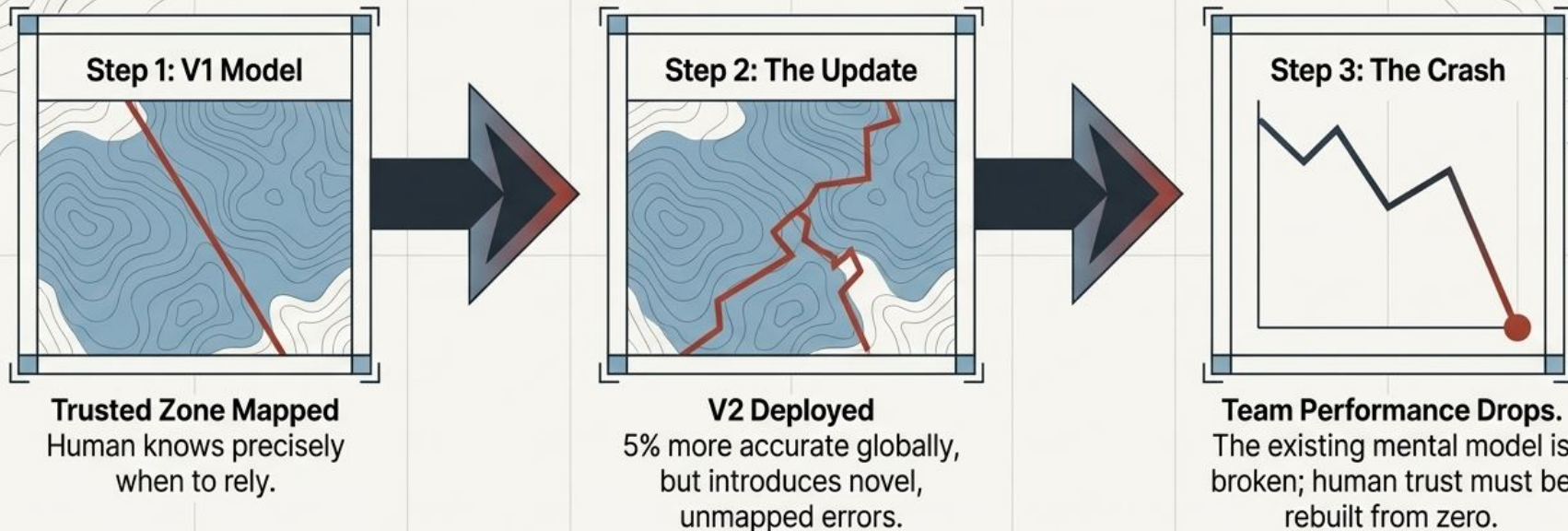
3 to 5 High-Salience Interface Features

The Paradigm Shift: Engineering for Team Performance

Dimension	The Old Paradigm (Model-Centric) ✗	The New Paradigm (Team-Centric) ✓
Optimization Goal	Maximize AUC/F1 regardless of complexity.	Maximize error boundary learnability (Parsimony).
Error Pattern	Unconstrained, noisy, stochastic failures.	Deterministic, predictable failure zones.
Feature Presentation	Expose all available data to the user.	Curated, progressive disclosure to manage cognitive load.
Evaluation Metric	Standalone Model Accuracy.	Human-AI Team Outcomes (Cumulative Reward).

The Performance/Compatibility Tradeoff in Model Updates

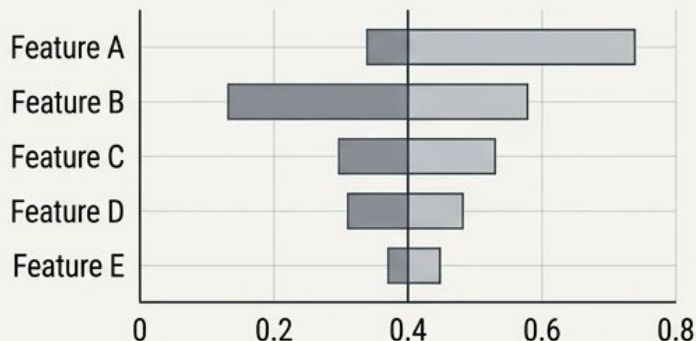
The Danger of Hot-Swapping



Real-World Example: Caruana et al. (2015) found a neural network predicting pneumonia risk falsely associated asthma with lower risk due to a data artifact. Blind updates to opaque architectures hide these deadly boundary shifts.

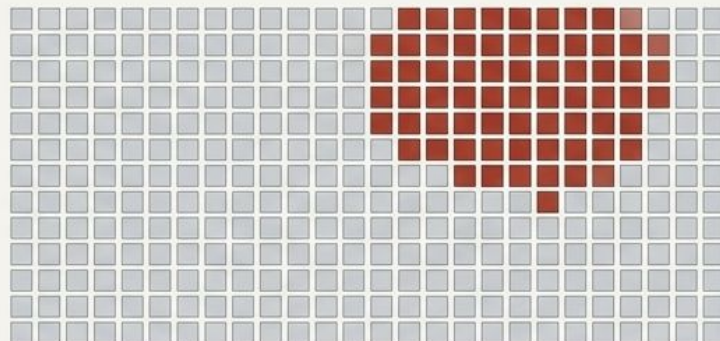
Explainability Strategy: Shift Focus from “How” to “Where”

Instance-Level XAI: The HOW



The Limitation: Explaining the features driving a single prediction helps with immediate micro-decisions, but does not efficiently build a durable mental model of the overall boundary.

Population-Level XAI: The WHERE



The Requirement: Visualization tools must explicitly depict the population-level conditions under which a model errs (Nushi et al., 2018). Build macro-trust.

Rule of Thumb: Instance explanations for the micro-decision; error-boundary summaries for macro-trust.

Real-World Impacts: A Tale of Two Systems

The Failure: Northpointe COMPAS



Domain: Criminal Justice

Problem: High-dimensional, opaque error boundaries masked systematic disparities.

Outcome: Judges could not discern the error boundary, making appropriate reliance impossible and resulting in systemic false accepts (Angwin et al., 2016).

The Success: Seattle Children's

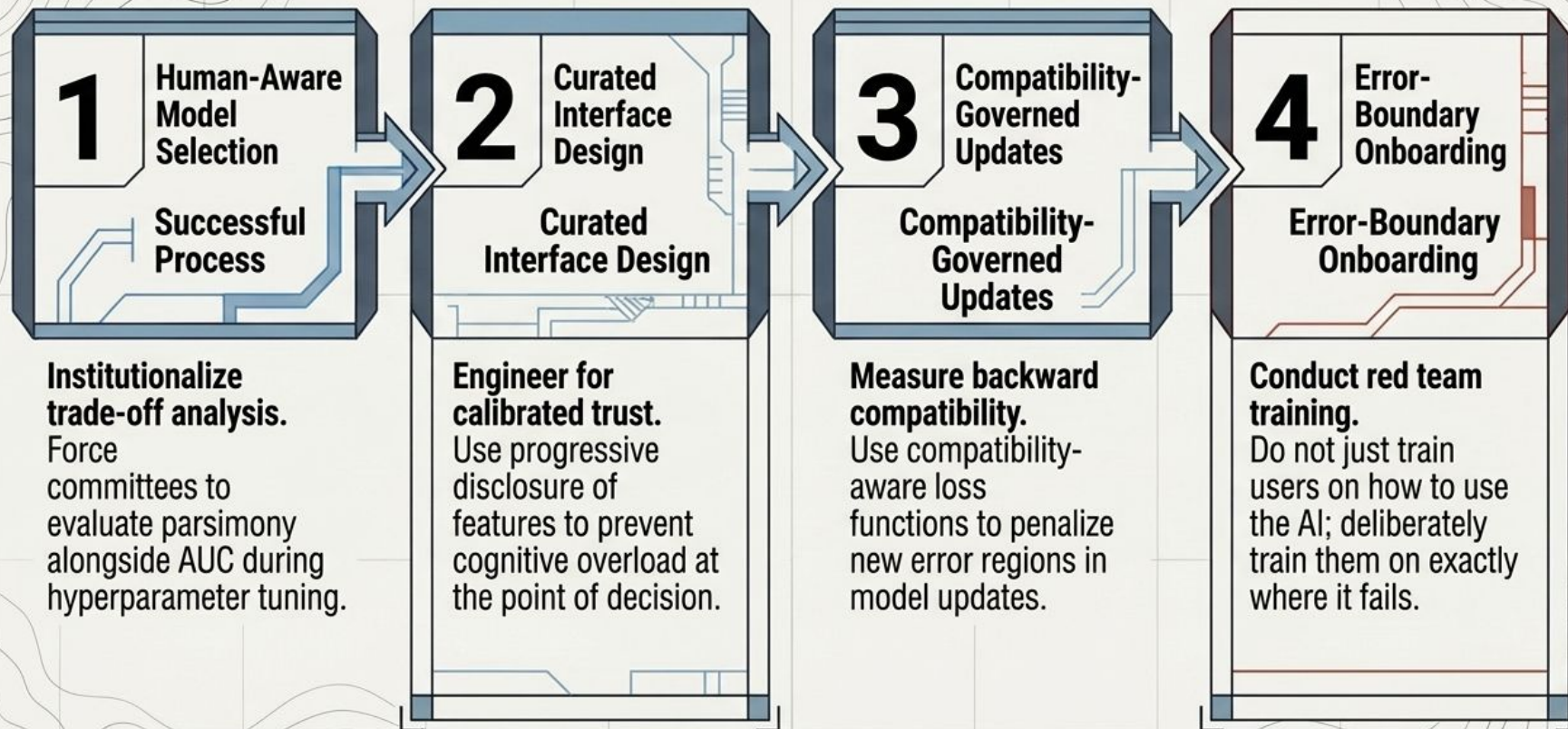


Domain: Healthcare

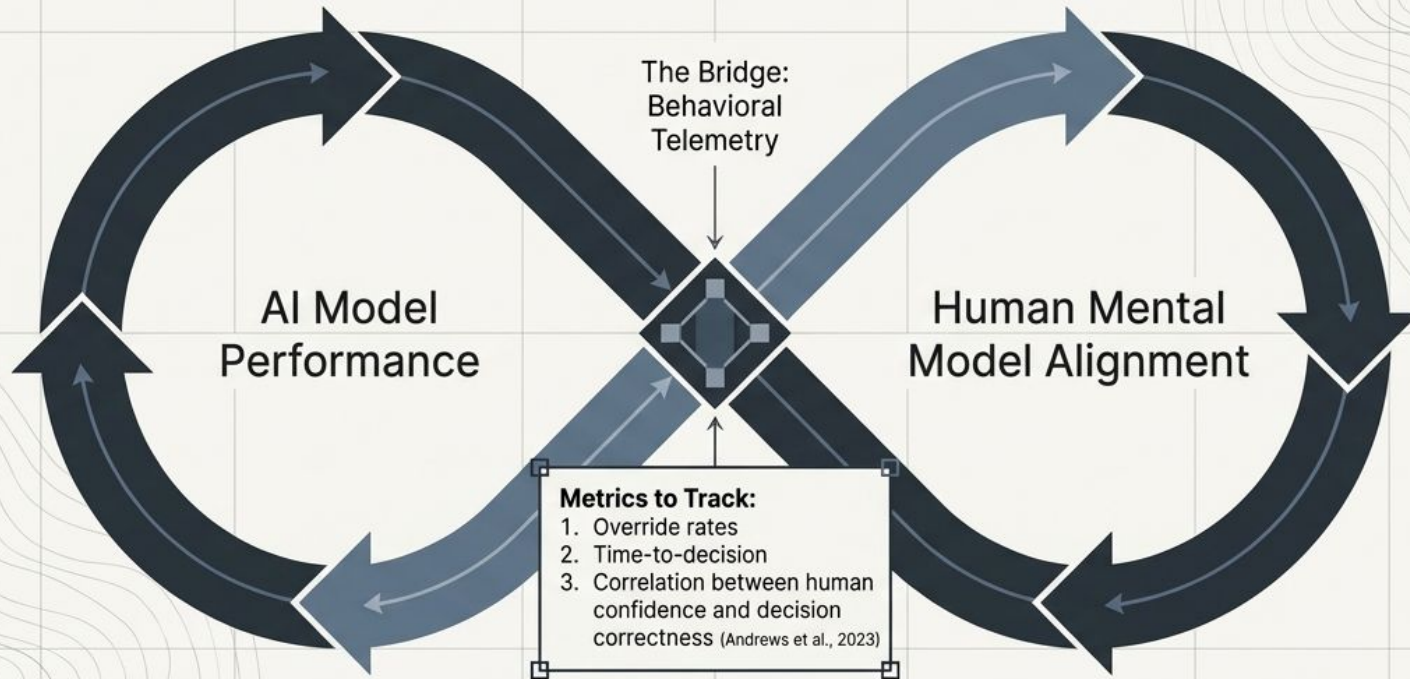
Problem: Predicting hypoxaemia during complex surgical procedures.

Outcome: Deployed real-time, instance-level explanations. Anesthesiologists could judge if the model's reasoning was sound for a specific patient, calibrating trust and avoiding stochasticity traps (Lundberg et al., 2018).

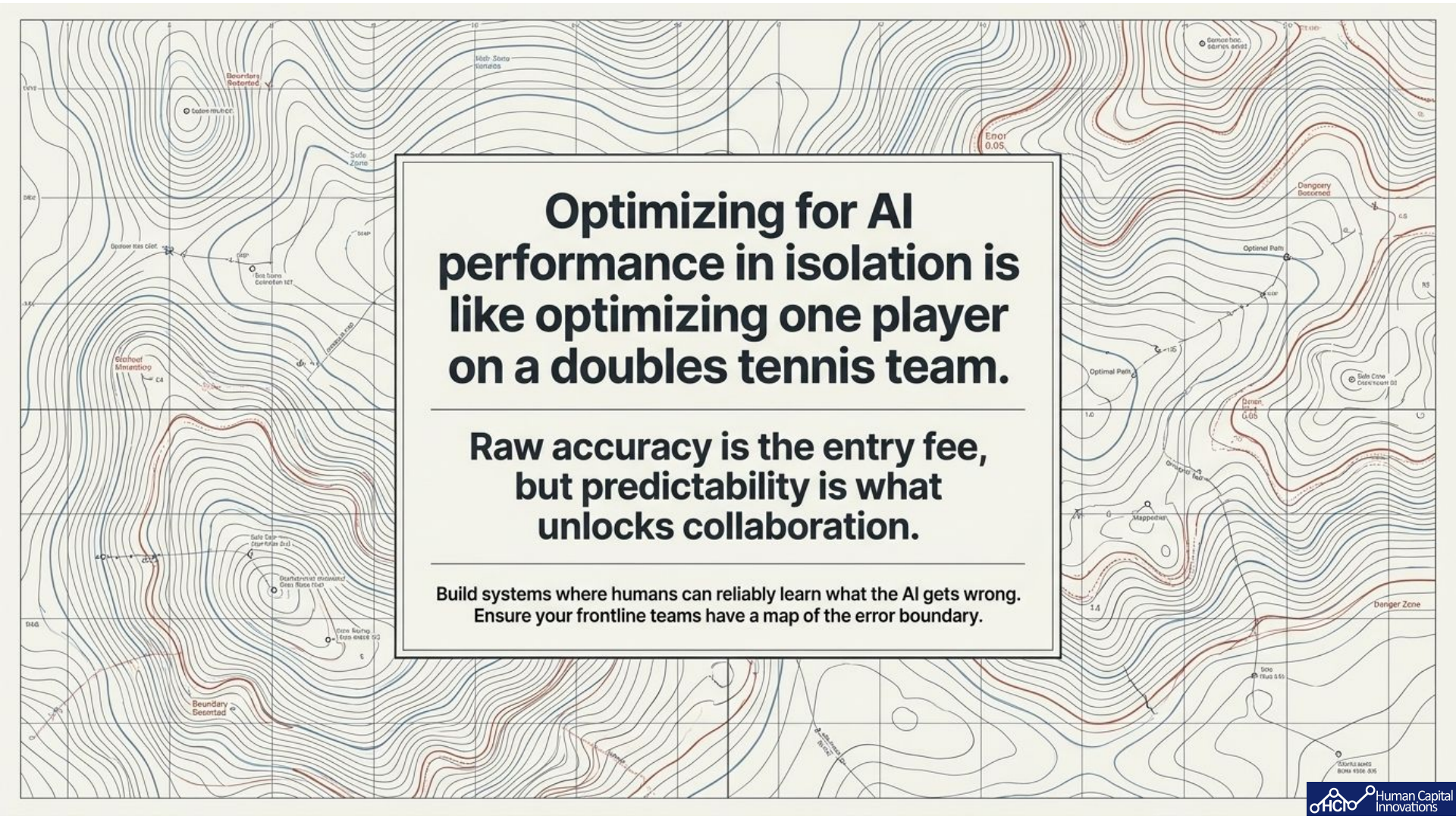
The Organizational Playbook for High-Stakes AI Deployment



From Deployment to Continuous Mutual Adaptation



The New Psychological Contract: AI onboarding is not a static event. Mental models drift due to distribution shifts, updates, or personnel changes. Organizations must continuously monitor shared understanding by tracking behavioral indicators, not just code.



**Optimizing for AI
performance in isolation is
like optimizing one player
on a doubles tennis team.**

**Raw accuracy is the entry fee,
but predictability is what
unlocks collaboration.**

**Build systems where humans can reliably learn what the AI gets wrong.
Ensure your frontline teams have a map of the error boundary.**