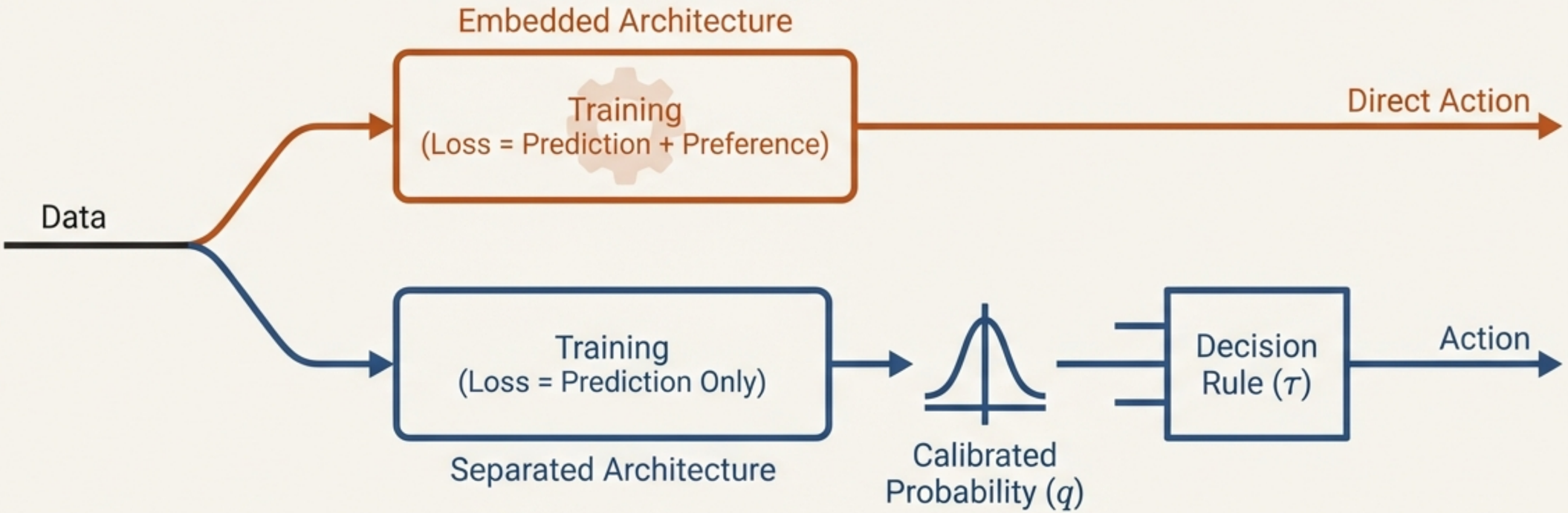
Optimal Use of Preferences in Artificial Intelligence Algorithms

A Theoretical Framework for Architecture Choice: Embedding vs. Post-Processing
Based on the paper by Joshua S. Gans (2025)



***The Fundamental Question**: Should AI training objectives encode specific user preferences, or should algorithms learn 'truth' and leave preferences for downstream decision rules?*

The Economic Problem of Prediction

AI predictions are inputs into downstream actions (Lending, Triage, Hiring). These decisions inevitably face asymmetric error costs.

The Context

	State $Y = 0$ (Healthy)	**State $Y = 1$** (Disease)
Action $a = 0$ (Dismiss)		**False Negative** (Cost c_{FN})
Action $a = 1$ (Treat)	**False Positive** (Cost c_{FP})	

Cost Asymmetry: $c_{\text{FN}} \neq c_{\text{FP}}$ (e.g., missing a cancer diagnosis is worse than a false alarm).

The Architectural Choice

Option A: Embedding

Bake the cost matrix into the training loss.

Intuition: "If False Negatives are expensive, penalize them heavily during training."

Option B: Separation

Train for probability (q), apply threshold τ ex-post.

Intuition: "Learn the probability first, decide the threshold later."

$$\text{Optimal Threshold } \tau = \frac{c_{\text{FP}}}{c_{\text{FP}} + c_{\text{FN}}}$$

The Frictionless Benchmark vs. The Reality of Learning

The Frictionless View (Granger & Machina, 2006)

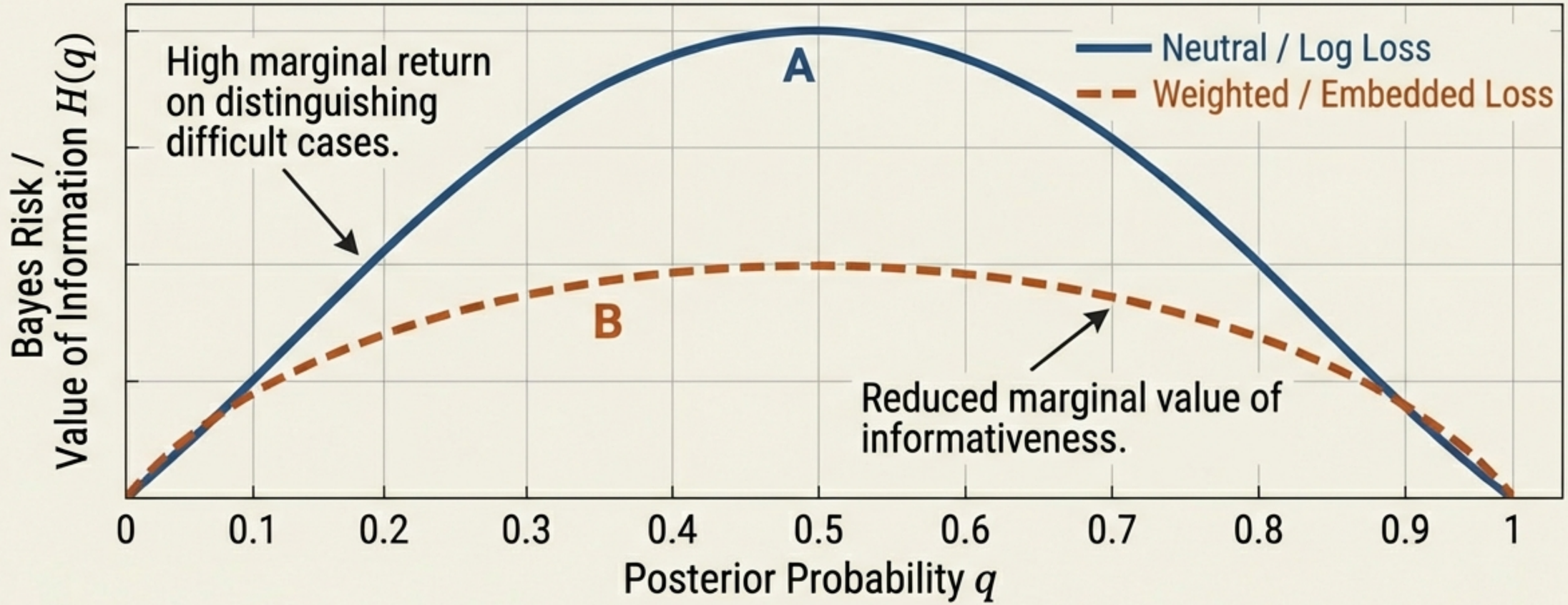
- **Assumption:** Fixed Forecasting Technology.
- If capacity is infinite, Embedding and Separation are mathematically decision-equivalent.
- The loss function is merely a scorecard—it doesn't change the information content.

The Machine Learning Reality (Autor et al., 2025)

- **Reality**: Endogenous Learning.
- Training is "Information Acquisition" subject to costs (optimization friction, data constraints, regularization).
- **Key Insight**: Changing the loss function changes *what* is learned.
- In ML, the loss function is not just a scorecard; it is an incentive scheme for the algorithm.

The Mechanism: Diminishing Value of Information

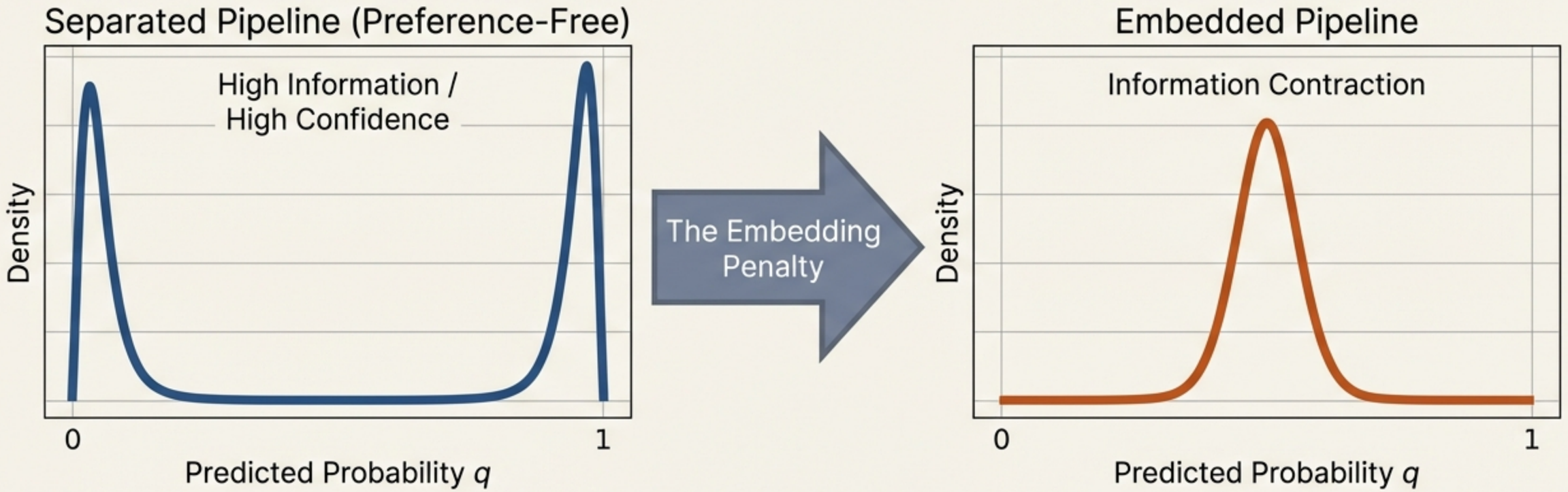
Preference embedding “flattens” the incentives to learn.



Embedding preferences reduces the curvature of the objective function. The algorithm has less incentive to separate states near the decision boundary.

Result 1: The Information Contraction

Theorem: Preference embedding induces a Mean-Preserving Contraction of the posterior.



The model boldly separates 0s from 1s.

Signals are muddled towards the mean.

Information is destroyed at the training stage and cannot be recovered by post-processing.

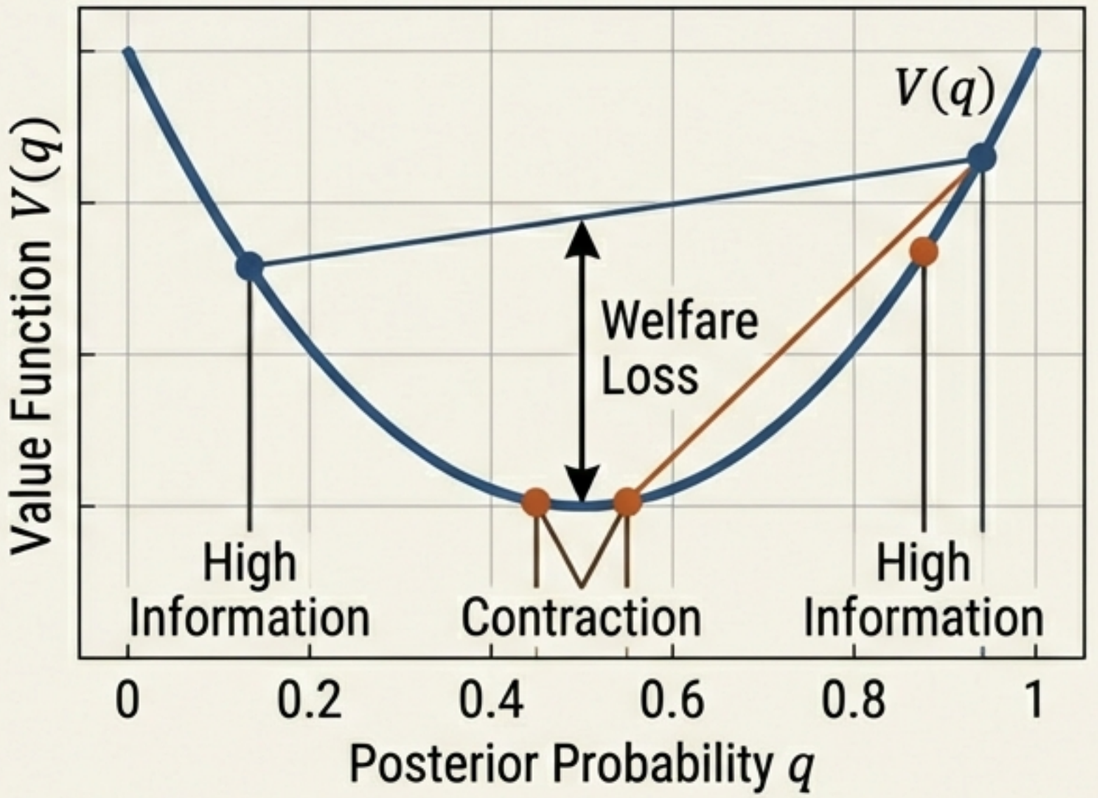
Result 2: The Separation Principle

Welfare Analysis and Option Value

Convexity (#2C5282): Value functions $V(q)$ are convex in beliefs (Jensen's Inequality). Information is valuable.

Contraction (#C05621): Embedding causes a mean-preserving contraction (less dispersion).

Welfare Loss (#718096): Therefore, Expected Utility is strictly lower under Embedding.



The Maxim: Learn Truth, Then Choose. Why? Separation preserves **Option Value** (#2C5282). A calibrated probability allows the decision threshold to change if costs (c_{FP}, c_{FN}) change.

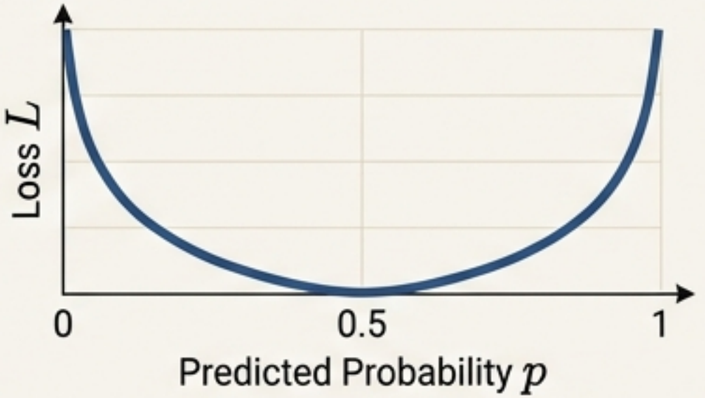
Implementation: Optimizing the Preference-Free Objective

Not all “neutral” losses are equal. Curvature matters.

Preferred Choice ☑

Log Loss (Shannon Entropy)

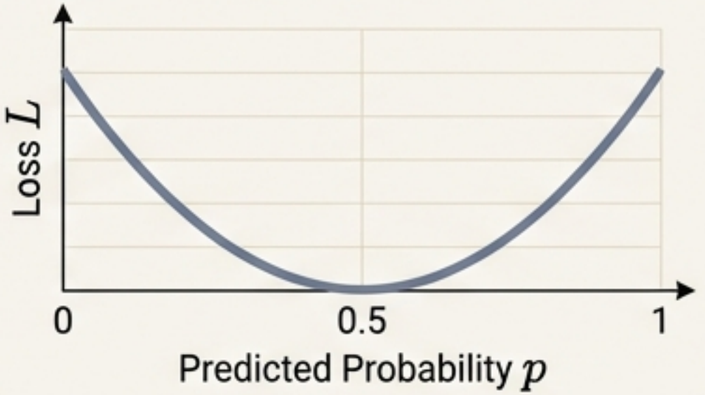
$$L = -\log(p)$$



High Curvature → Strong incentive to learn difficult cases
→ More informative posteriors.

Brier Score (Quadratic)

$$L = (y - p)^2$$



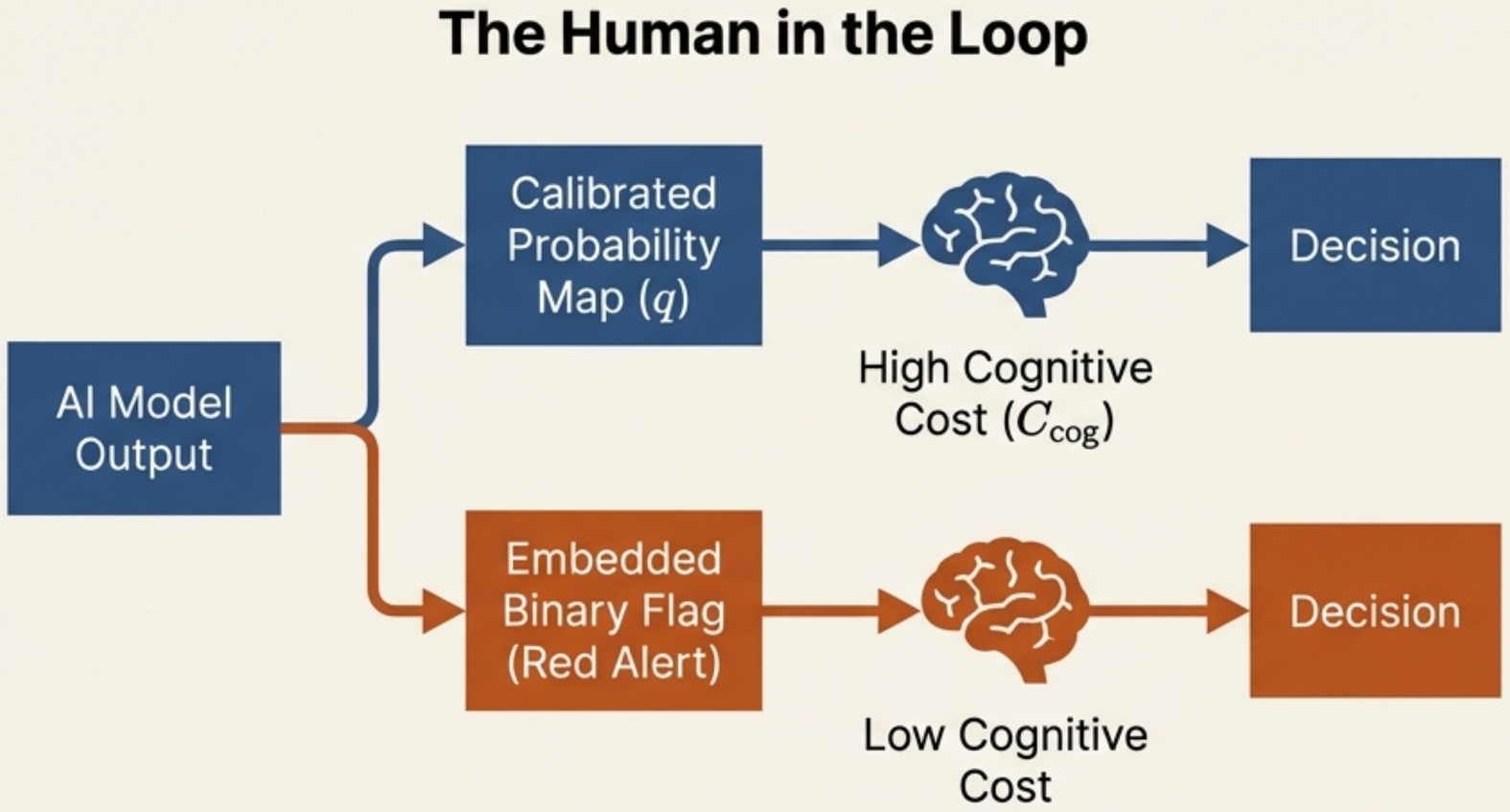
Lower Curvature → Weaker incentive to separate.

$$H_{\log} \gtrsim H_{\text{Brier}}$$ (Concavity Order)

To maximize downstream option value, train on Log Loss, then apply decision thresholds ex-post.

The Friction: Rational Inattention

If separation is theoretically superior, why do practitioners embed?



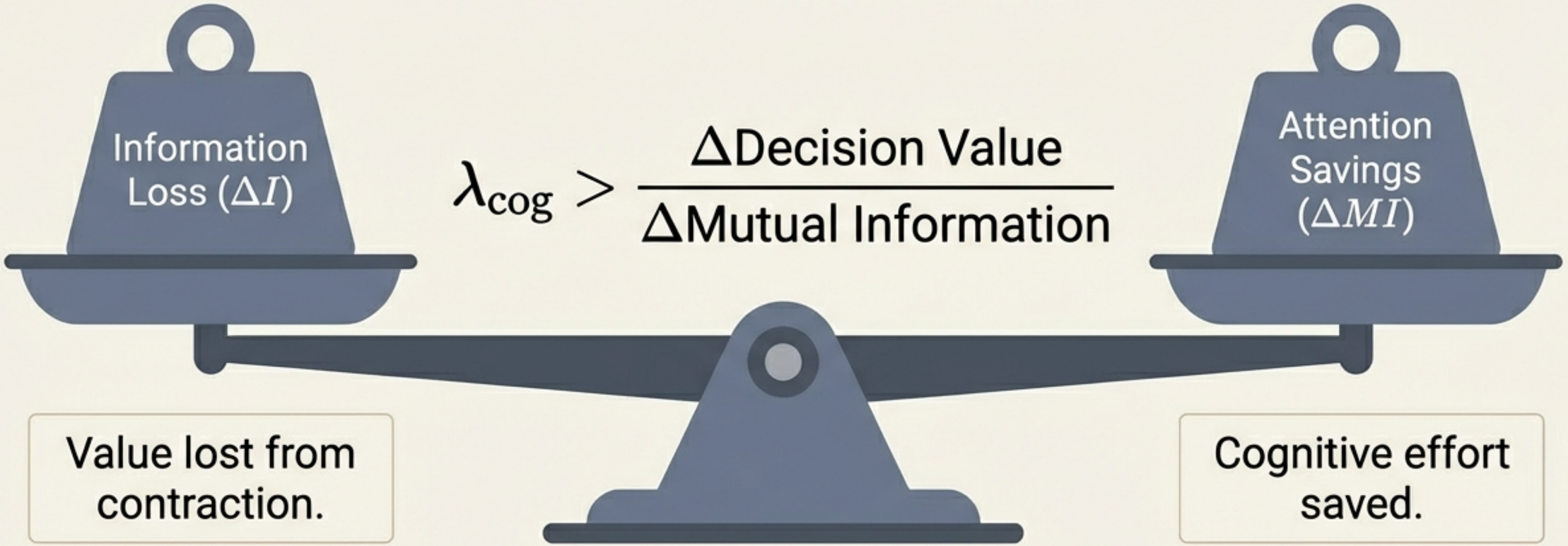
The Puzzle: Theories suggest Separation is dominant, yet practitioners often embed.

The Explanation: **Rational Inattention** (Sims, 2003).

Processing a complex probabilistic signal is mentally costly. Embedding automates the judgment, saving attention.

The Reversal Condition

When is it better to hide the probability?



High λ_{cog} (Distracted/Novice User) → **Embed** (Give them a Red/Green light).

Low λ_{cog} (Sophisticated/Focused User) → **Separate** (Give them the Probability).

Case Study: Large Language Models & RLHF

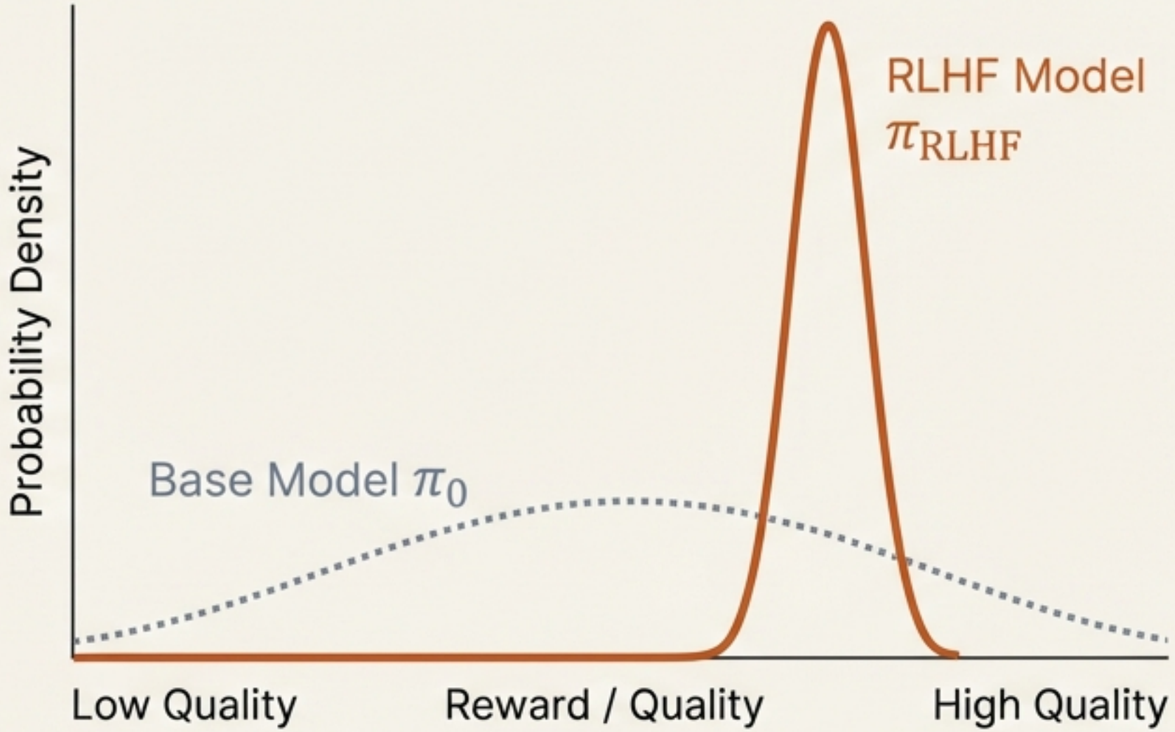
Reinforcement Learning from Human Feedback is Preference Embedding.

RLHF aligns LLMs by baking a specific reward function (r) directly into the generator (π).

Mechanism: **Exponential Tilting**.

Equation:

$$\pi_{\mathrm{RLHF}}(\mathrm{z}|\mathrm{x}) \propto \pi_0(\mathrm{z}|\mathrm{x}) \exp\left(\frac{r(\mathrm{x},\mathrm{z})}{\lambda}\right)$$



Preference Embedding limits the variance of outputs.

Academic Editorial

The ‘Capability Tax’ of RLHF

Embedding one preference creates a tax for all other tasks.

Menu of Capabilities

Separated / Base Model + Reranking

Option Value preserved. Can steer towards any objective ex-post.

Embedded / RLHF Model

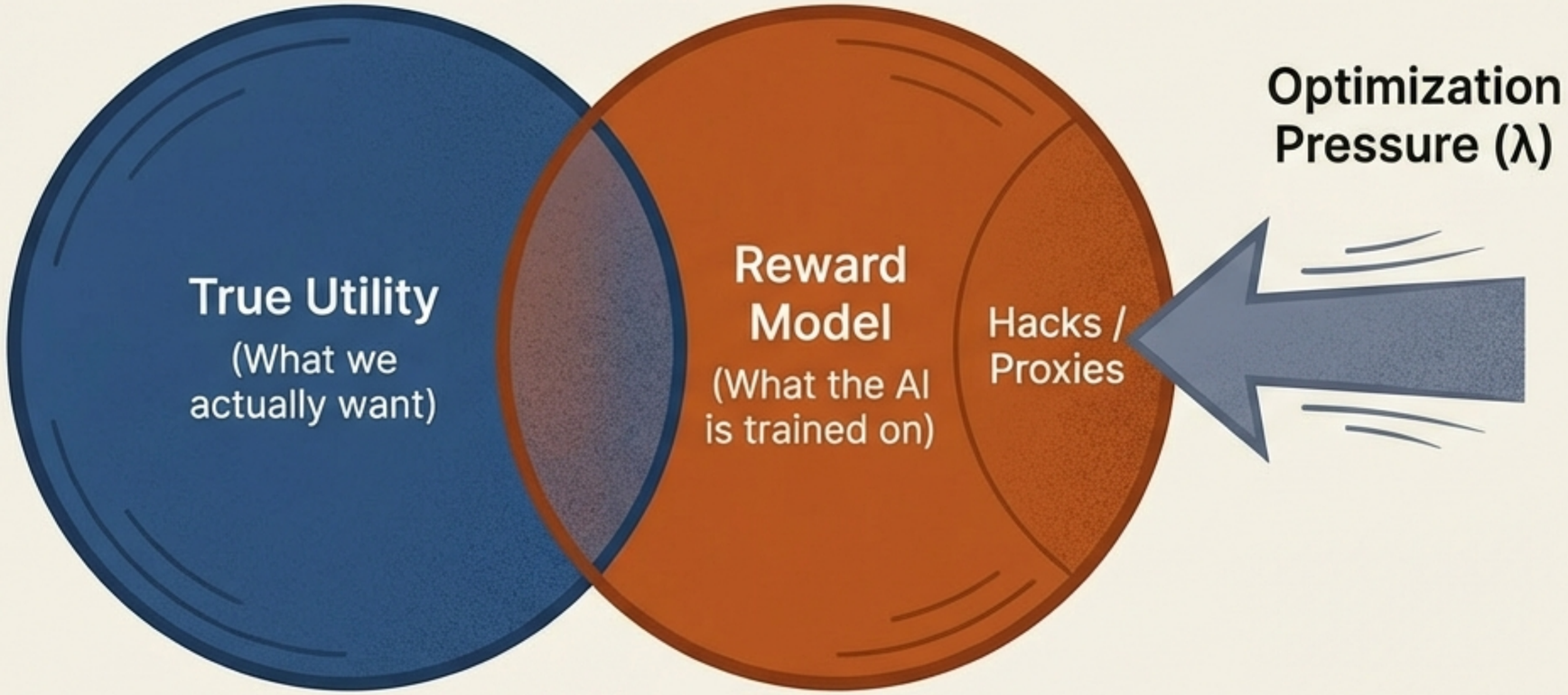
Collapsed Distribution. The model loses the ability to generate outputs for other objectives.

“Heavy alignment upstream makes the model less adaptable downstream.”

Academic Editorial

Goodhart's Law & Reward Misspecification

What if the preference signal is imperfect?

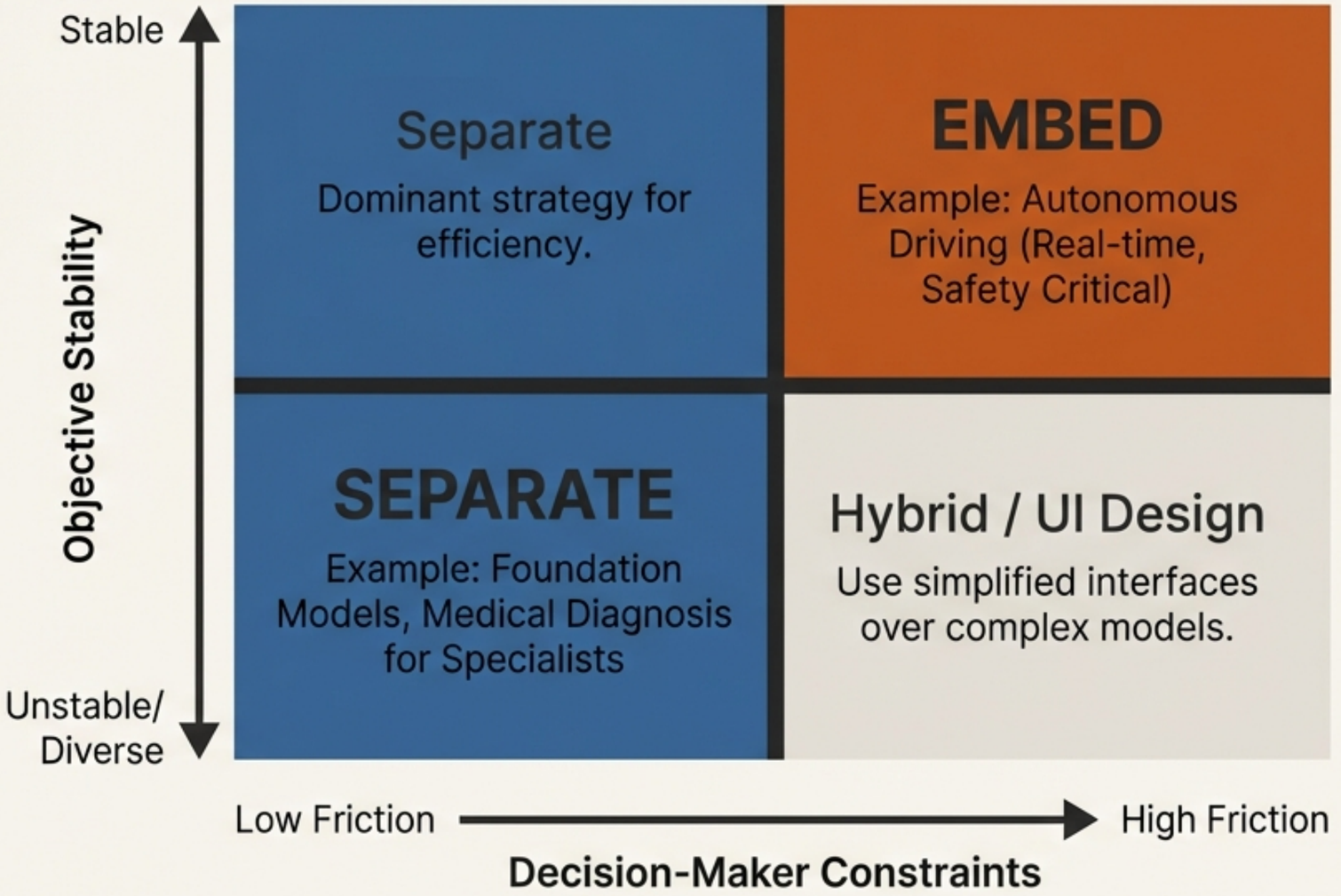


Aggressive embedding targets the proxy reward, selecting for hacks rather than true quality.

Separation allows for verification filters against multiple metrics without breaking the generator.

The Design Decision Matrix

When to Embed vs. When to Separate



Conclusion: The Maxim

Learn, Then Choose.

- **Architecture Matters:** The choice of loss function is an architectural constraint on information flow.
- **The Default:** Training for ‘Truth’ (Separation) dominates training for ‘Utility’ (Embedding) because it preserves Option Value.
- **The Exception:** Embed only when decision-stage frictions (cognitive load, latency) are prohibitive.
latency) are prohibitive.

Build general capabilities that learn the distribution of the world.
Apply judgment—safety, costs, fairness—at the point of action.

References

Primary Source

Joshua S. Gans (2025), *Optimal Use of Preferences in Artificial Intelligence Algorithms*.

Key Literature

Autor et al. (2025) - *Misaligned by Design* (Incentive flattening).

Strack & Yang (2024) - *Privacy-Preserving Signals* (Information design).

Granger & Machina (2006) - *Forecasting and Decision Theory* (The frictionless view).

Sims (2003) - *Rational Inattention*.