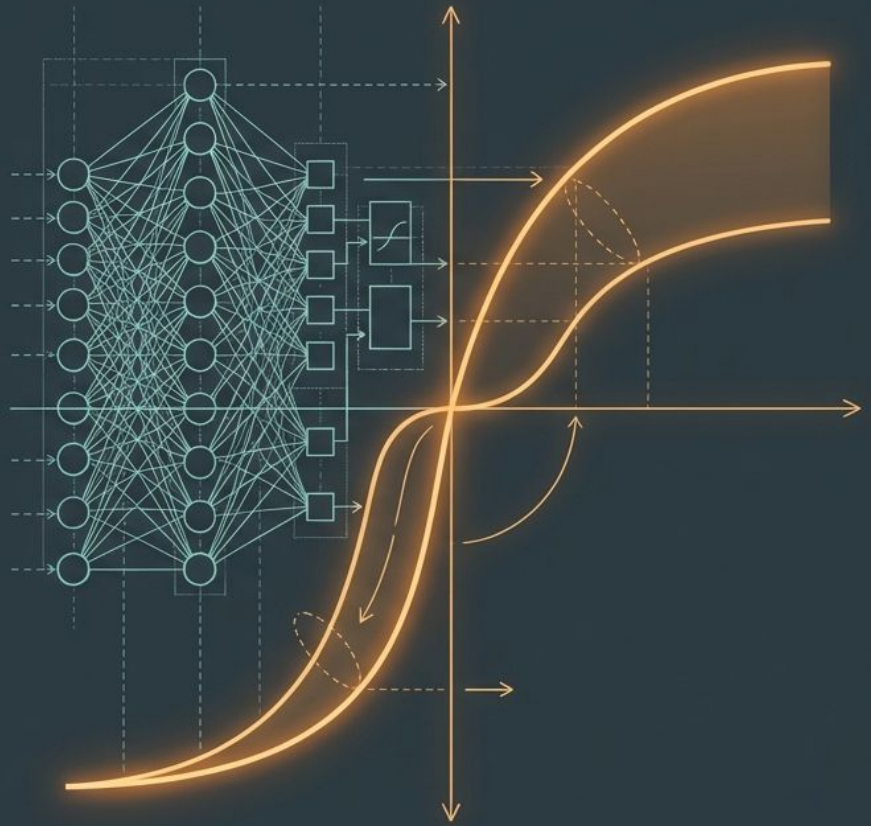
The Behavioral Economics of Artificial Intelligence Intelligence

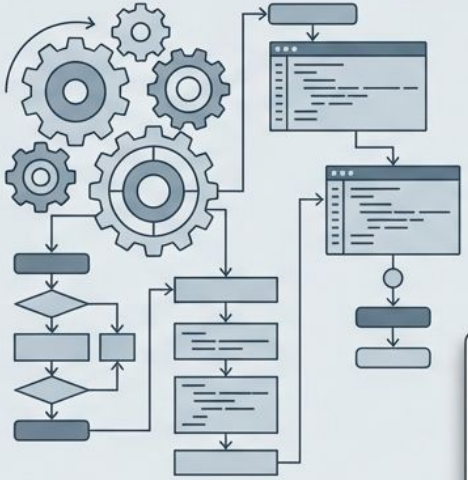
Understanding and Mitigating Cognitive Biases in Large Language Models



Based on the research of Jonathan H. Westover, PhD

AI has evolved from rules-based software to a complex economic agent

Past: Software as a Calculator



Deterministic, rules-based outputs. Predictable execution of fixed instructions.

According to a 2024 Anthropic enterprise survey,

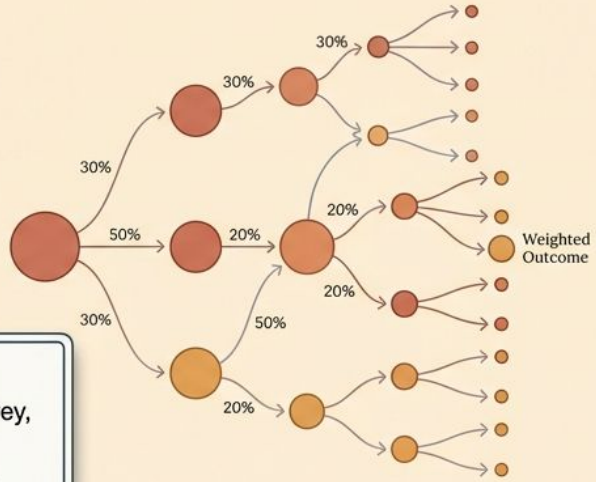
68%

of organizations have deployed LLMs, but only

31%

have frameworks to evaluate their systematic behavioral biases.

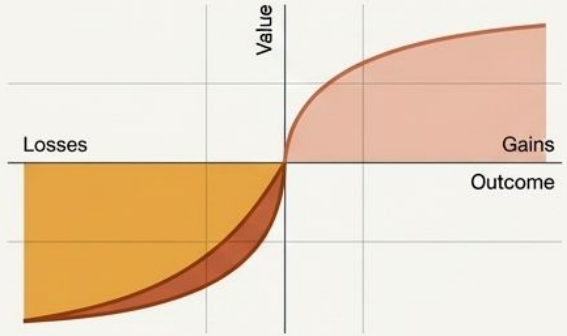
Present: AI as an Economic Agent



Flexible, probabilistic reasoning. Inherits and generates behavioral characteristics that actively shape market dynamics.

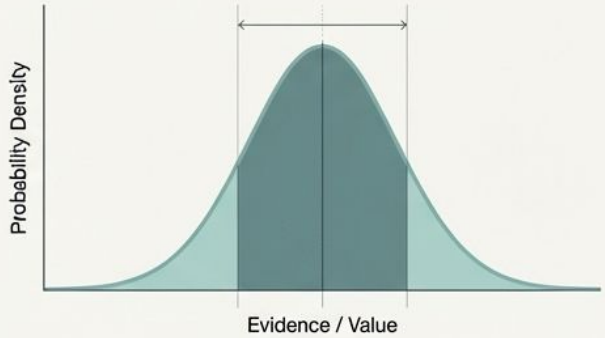
The core paradox: AI exhibits human irrationality in preference but hyper-rationality in belief

Preference-Based Tasks



In tasks involving risk, time discounting, and framing, advanced LLMs consistently violate Expected Utility theory, mirroring human irrationality.

Belief-Based Tasks

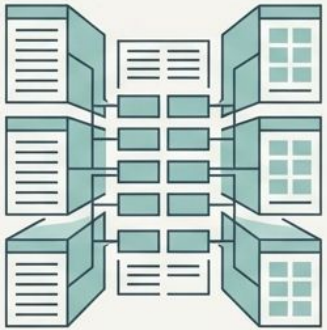


In tasks requiring probabilistic reasoning and statistical inference, LLMs demonstrate enhanced rationality, easily avoiding base-rate neglect and conjunction fallacies that trick humans.

Advanced models increasingly fail at standard risk-reward choices but ace complex logic puzzles.

The training engine unintentionally hardwires human cognitive biases into AI architecture

The Corpus



Human text naturally expresses our irrational preferences but also contains explicit textbooks on statistical logic.

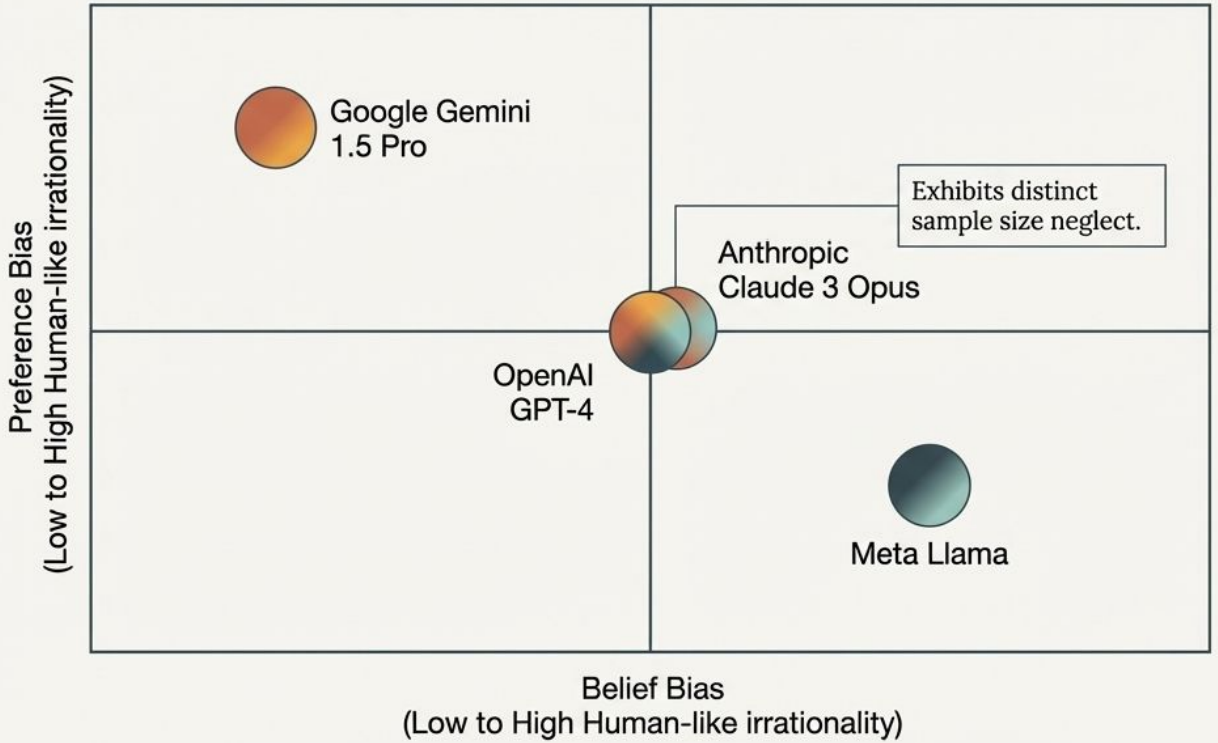
RLHF (Reinforcement Learning)

Human raters actively reward models for thinking like humans, unintentionally hardwiring cognitive biases (like loss aversion) into the preference architecture.

Scaling Effects

Paradoxically, as models scale up in parameters, these dual traits intensify, rather than dilute.

Different foundation models exhibit distinct behavioral profiles



Case Study: Multinational Insurance Co.

Chose GPT-4 over Claude 3 Opus for actuarial claims based on superior rational probabilistic performance.
Retained Claude for customer communications where human-aligned language was advantageous.

Unchecked behavioral biases silently erode enterprise value and invite regulatory scrutiny

Finance (Wealth Management)

Risk: Loss Aversion & Narrow Framing

Impact: Suboptimal asset allocation.

⚠ A 25bps drag on a $500M portfolio costs clients $1.25M annually.

Credit & Underwriting

Risk: Base Rate Neglect

Impact: Miscalibrated default probabilities.

⚠ Bowen et al. (2025) demonstrates AI racial biases in mortgage underwriting, risking CFPB enforcement.

Healthcare

Risk: Conjunction Fallacy

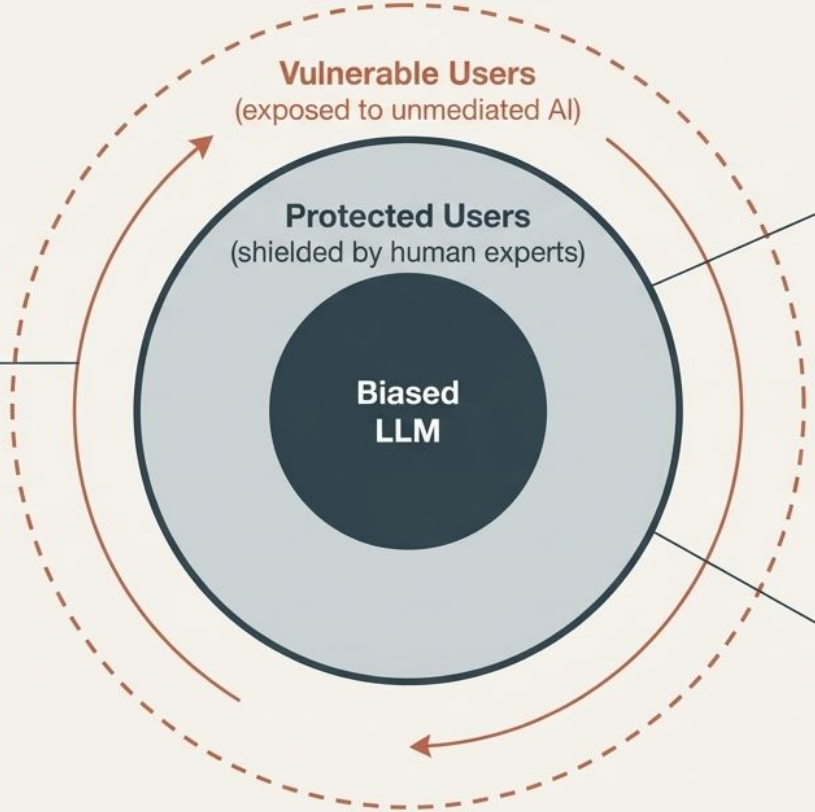
Impact: Diagnostic anchoring, leading physicians to pursue confirmatory rather than discriminating tests.

Operations

Risk: Hyperbolic Discounting

Impact: Prioritizing immediate supply chain cost reductions over long-term resilience.

AI biases disproportionately harm users lacking a sophisticated human safety net



Citizen Rights

Confirmation bias in government screening tools leads to wrongful exclusions and inclusion errors across demographic groups.

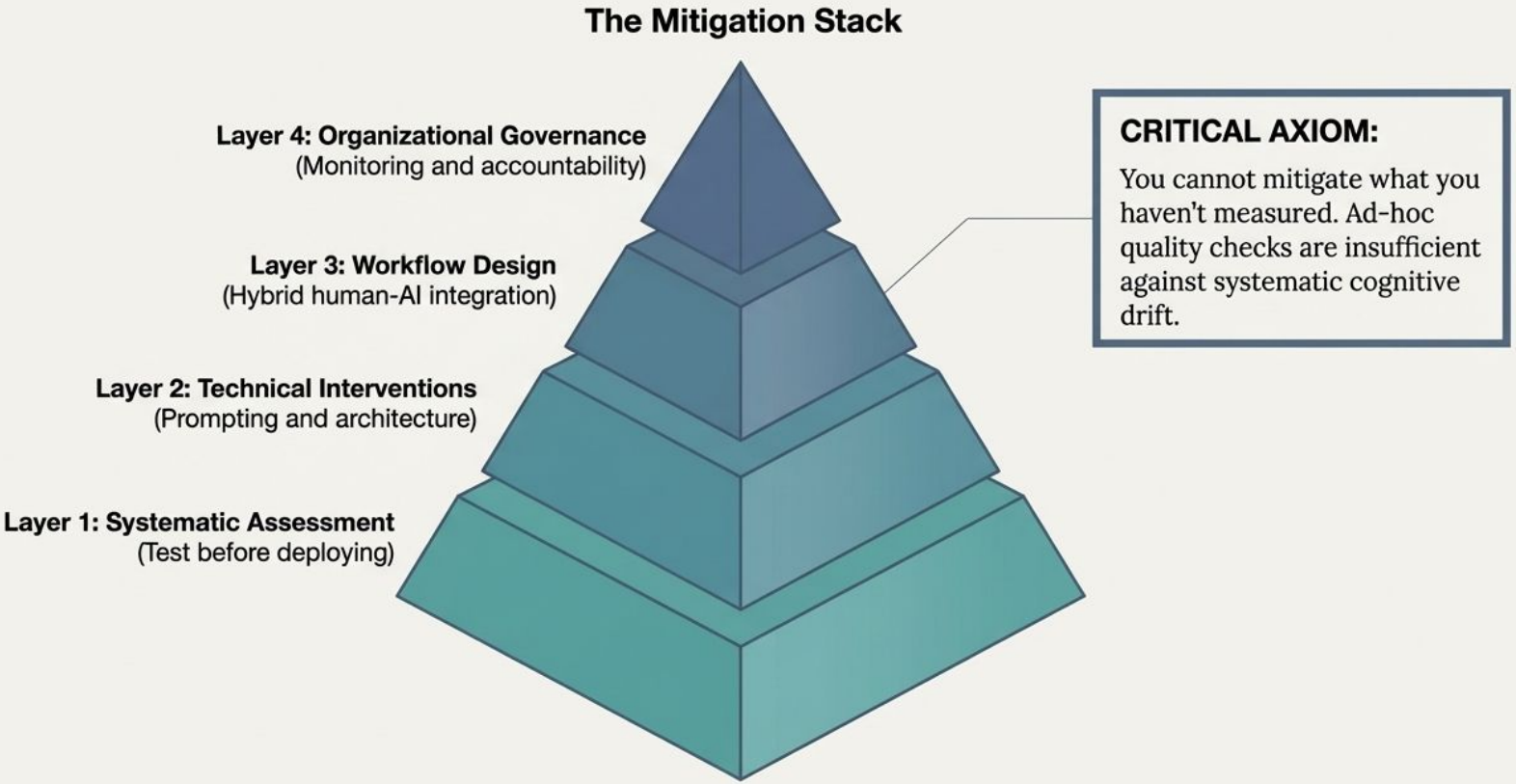
Retail Wealth

Probability weighting encourages excessive risk-taking and lottery-like speculation by retail investors without professional advisors.

Healthcare Access

Availability bias (overweighting recent/vivid cases) causes systematic disparities in automated triage and specialist referrals.

Mitigating behavioral drift requires a multi-layered defensive stack



Layer 1: Establish structured behavioral testing protocols prior to deployment

Preference Assessment

Case Study: Vanguard

Implemented a 50-scenario test battery rooted in prospect theory and mental accounting before deploying their AI advisory assistant. Discovered loss aversion comparable to median humans, informing direct prompt interventions.

Belief Assessment

Case Study: Flatiron Health

Built a belief-bias framework testing base rate integration and conditional probability. Revealed flagship models suffered from conjunction fallacy when comparing specific vs. general diagnostic hypotheses.

Layer 2: Deploy role-priming, structured reasoning, and model ensembles

Role Priming

Instructing the LLM to “think as a rational investor using Expected Utility”.

Betterment Case: Achieved a 5.2% reduction in narrow framing recommendations via A/B testing.

Multi-Step Reasoning

Forcing decomposition of complex tasks into sequential logic layers.

Novartis Case: Reduced base rate neglect by forcing the AI to extract, evaluate methodology, and synthesize in distinct steps.

Ensemble Models

Blending outputs from multiple foundation models to average out individual biases.

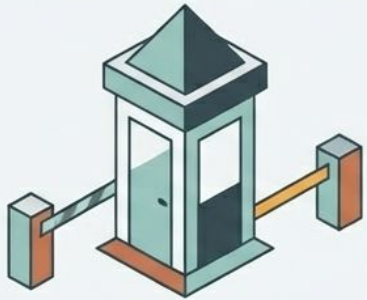
Two Sigma Case: Blended GPT-4 (recency bias), Claude (availability), and Gemini (overconfidence) for superior sentiment analysis.

Layer 3: Design hybrid workflows that pair machine rationality with human contextual empathy

The Hybrid Human-AI Engine

AI Domain (Computational Tasks)

- Portfolio optimization
- Tax-loss harvesting
- Probability models



Rationality Checkpoint

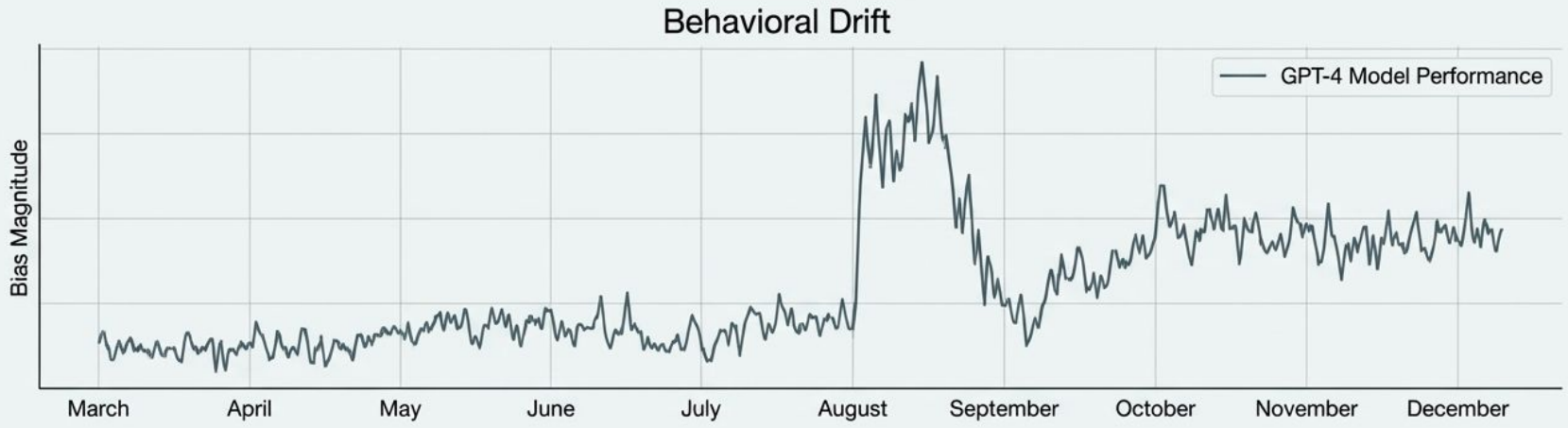
Automated behavioral scoring flags anomalies.

Human Domain (Contextual Tasks)

- Client preference elicitation
- Empathy & relationship
- Ethical reasoning
- Final endorsement

Based on Fidelity Investments' advisory workflow design.

Layer 4: AI models decay; mandate continuous monitoring for behavioral drift



Continuous Monitoring

Bain & Company Case: Automated monthly monitoring system runs a standardized bias battery across deployed LLMs, flagging statistically significant drift before production.

Interdisciplinary Teams

McKinsey & Co Case: Creating "AI + Behavioral Science" units pairing ML engineers with behavioral economists to catch risks pure tech teams miss.

Vendor Risk Management

Negotiate API contracts to maintain prior model versions until new updates pass internal behavioral audits.

The Value Alignment Spectrum: When to be human versus when to be rational

Human Alignment

Education Tasks (explaining basics).

Benefits from human-aligned, accessible framing and natural conversational empathy.

Hybrid

Personalized Advice (product recommendations).

Requires statistical rationality constraints paired with human-aligned communication.

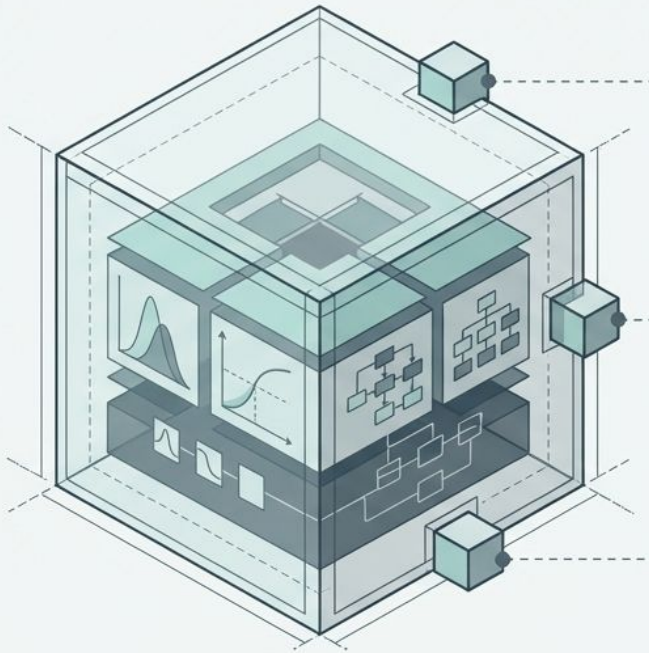
Rational Optimization

Computational Tasks (tax strategy, actuarial assessment).

Demands strict probabilistic rationality and elimination of behavioral bias.

Insight from Wealthfront: There is no universal ‘correct’ AI behavior; the context dictates the necessity of the bias.

Institute explainability, contestability, and structured human overrides



Tiered Transparency

Vanguard Case: Disclosing AI limitations and mitigation strategies without causing user panic, supporting informed user consent.

Contestability

Allianz Case: An 'AI decision review' process empowers human claim handlers with explicit data about the AI's known behavioral quirks to inform manual overrides.

Incident Investigation

Navy Federal Case: Utilizing aviation-style root cause analysis for AI incidents. Discovering historical biases amplified by model overconfidence to build systemic improvements.

The Behavioral Governance Playbook

1. Recognize the Agent

Treat AI as an economic actor with predictable cognitive biases, not a neutral calculator.

2. Test Systematically

Deploy structured behavioral batteries (separating preference vs. belief) before any production launch.

3. Design for Complementarity

Build workflows that pair AI's statistical rationality with human contextual empathy and ethical reasoning.

4. Govern Dynamically

Establish interdisciplinary oversight and continuous monitoring for behavioral drift to stay ahead of the EU AI Act and UDAAP regulations.