AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

KRISH VASUDEVAN
Founder, TERAXYS LLC

TERAXYS LLC · AUGUST 2026

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

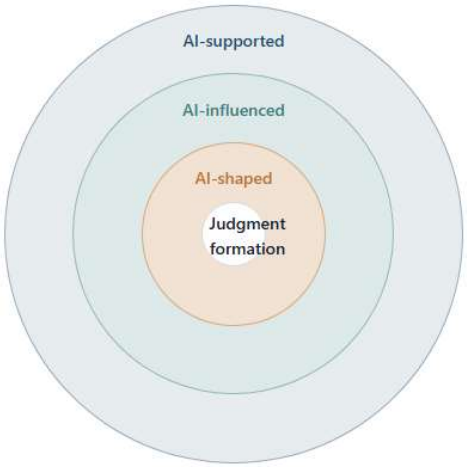
I. The Hidden Architecture of Judgment

A consequential decision may be formally approved while revealing almost nothing about how the judgment beneath it was actually formed.

AI can enter that process at different levels. It may supply information to a judgment that remains substantially independent, materially alter the decision-maker's reasoning, or construct much of the analytical frame within which the decision is ultimately made. These are distinct cognitive configurations, with different economic benefits and exposures.

FIGURE 1

Configurations of AI involvement in judgment formation



01 **AI-supported**
Supplies evidence, research, calculations or synthesis to a substantially independent judgment.

02 **AI-influenced**
Materially changes interpretation, assumptions or the alternatives considered.

03 **AI-shaped**
Constructs much of the frame, analytical architecture or recommended conclusion.

Increasing depth of influence over judgment formation

The layers are nested because earlier forms of participation can remain present as AI moves deeper into judgment formation. Depth represents influence—not error, severity or a risk ladder.

The stacking reflects influence. As AI moves deeper into judgment formation, earlier forms of participation do not necessarily disappear. AI-shaped judgment may still rest on AI-supplied evidence and AI-influenced interpretation. What changes is the extent to which the model participates in constructing the judgment itself.

AI-supported judgment

In **AI-supported judgment**, the decision-maker remains the principal source of the frame, analysis and conclusion. AI supplies evidence, research, calculations, synthesis or other analytical inputs.

Its apparently limited role can obscure an upstream vulnerability: before difficult analysis begins, AI may identify the wrong source, misunderstand a credible one, omit a material fact or supply a plausible interpretation that alters the decision-maker's premise. We then face a systemic crisis of *fontes ipsi venenantur* – the very sources are poisoned. The decision-maker is no longer relying on competent analysis, but navigating the synthetic byproduct of a corrupted foundation.

AI support → accepted premise → independent analysis → judgment

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

Framing can fail earlier still. The question may be defined poorly before the model is engaged, or the model may understand a reasonable frame differently from what was intended. Sophisticated analysis can then develop around a mistaken conception of the problem. A capable model can misread a difficult problem intelligently, and an equally capable decision-maker can articulate a complex problem poorly or further reason intelligently from that misreading.

AI-influenced judgment

In **AI-influenced judgment**, AI materially changes part of the reasoning while substantive independent judgment still remains active. The model may expose contradictory evidence, challenge an assumption, introduce an alternative or cause the decision-maker to reinterpret a relationship.

This may be among the most economically valuable configurations available from current AI. Decision-maker and model can interrogate the same problem differently, exposing weaknesses in one another's reasoning and producing a conclusion superior to either working independently.

The same interaction permits **shared delusion**.

Two cognitive participants do not necessarily provide two independent checks. Decision-maker and AI can converge on the same sophisticated but incorrect interpretation. The model can then generate further analysis consistent with the shared premise, which may be read as corroboration when it merely validates the premise already accepted.

Challenge can become reinforcement with no explicit reference point that the transition has occurred.

AI-influenced judgment is therefore not an intermediate point on a simple risk spectrum. It may have the widest range of outcomes: from exceptional augmentation at one end to mutually reinforced error at the other.

AI-shaped judgment

In **AI-shaped judgment**, the model materially determines the problem frame, principal assumptions, analytical architecture, alternatives considered or recommended conclusion. Formal authority may remain entirely with the decision-maker while the intellectual structure presented for consideration is substantially model-generated.

Here the limits of independent understanding begin to fray. AI can assemble and analyze information faster than a decision-maker can reconstruct it, extract sophisticated data across domains and combine bodies of evidence whose individual elements remain intelligible while their end-to-end construction becomes increasingly difficult to reproduce.

Review may remain visible in form while becoming thinner in substance. A capable reviewer may challenge assumptions, request additional analysis and approve the recommendation without independently reconstructing enough of the analytical architecture to challenge its governing frame or even correlated multi-domain assumptions.

human review ≠ independent human judgment

At the extreme, review can retain the appearance of scrutiny while becoming a **rubber stamp in substance**.

This need not reflect carelessness. Full independent reconstruction may simply become impractical at the speed, breadth or complexity of model-led analysis. Responsibility remains with the decision-maker even as dependence on AI increases for the intellectual structure through which the decision is understood.

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

The invisible part of the judgment

The organization may not reliably know which configuration produced a particular decision.

AI participation can be **unobserved**, where its contribution never becomes visible in the formal decision process, or **unauthorized**, where it occurs outside an explicit organizational boundary. One concerns visibility and the other authority, but both can understate AI's role in forming substantive judgment.

The finished product may reveal little. AI-assisted analysis can be an ordinary spreadsheet, financial model, memorandum or presentation, substantially edited and reviewed before approval. Nothing in it necessarily reveals how its underlying judgment was constructed.

The relevant question is therefore not simply whether AI was used, but its **end-to-end influence over the judgment that has entered the organization**: whether AI supported, influenced or shaped it; whether decision-maker and model operated within the same intended frame; and whether that participation occurred within the visible and authorized perimeter.

The ease with which judgment can move between these configurations depends partly on another characteristic of AI: it changes the time and cost of forming, testing and challenging an independent view.

II. The Value of Friction

The economic value of AI in judgment-intensive work begins with compression.

Work that once required a professional to construct an answer methodically can increasingly begin from an apparently appropriate default solution. AI therefore changes more than the speed of execution. It changes the point at which independent reasoning begins.

Evidence from complex professional work shows the attraction. In a preregistered experiment involving 758 Boston Consulting Group consultants, GPT-4 users completed 12.2 percent more tasks, worked 25.1 percent faster and produced outputs rated more than 40 percent higher in quality within the model's capability frontier, but were 19 percent less likely to solve a selected task outside it correctly.[1] Field experiments with 4,867 professional software developers found an approximately 26 percent increase in completed tasks with AI coding assistance.[2] In a randomized trial involving 92 physicians, GPT-4 assistance improved complex patient-management reasoning by 6.5 percentage points; the physicians also spent more time on the cases, indicating that AI altered the substance of reasoning rather than merely accelerating it.[3]

The economic gain is clear. The less obvious question is **what kind of work is being compressed**.

Some friction is wasteful: effort that consumes time without materially improving judgment. Other friction forces the decision-maker to develop enough independent understanding to test the problem rather than merely accept its apparent solution.

Reading a source instead of relying on its summary can expose context that changes its meaning. Reconstructing an assumption can reveal that it was never justified. Working through an unresolved contradiction can change the frame itself. Developing an alternative independently can expose a possibility that never appears once one plausible answer has become the default.

This is **judgment-forming friction**.

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

Its cost is immediately visible. Its value is often visible only when it prevents a poor judgment.

AI changes that balance by supplying a plausible starting position before the decision-maker has constructed one independently. Once that position exists, the task changes. Instead of building an analysis from the underlying problem, the decision-maker increasingly evaluates whether there is sufficient reason to depart from an analysis already available.

Time pressure magnifies the asymmetry. A meta-analysis covering 125 studies and 827 effect sizes found that time pressure generally increased speed while reducing accuracy across cognitive and perceptual tasks; other experimental research has associated constrained decision time with greater reliance on simplified processing.[4]

With AI available, pressure can also change the **sequence of judgment formation**.

When time permits, a decision-maker can form an independent view and use AI to test it. As available time contracts, AI can become the starting point instead. Its frame substitutes for the construction of an independent one. Its synthesis becomes the practical representation of the evidence because reopening the sources would consume the remaining time. Alternatives are then evaluated from within a structure already established by the model.

By the time formal review occurs, the question may no longer be *What do I think?* but *Is there enough reason to reject what is already here?*

That creates **escalating deference**.

> AI default → accepted starting frame → AI-mediated evidence → AI-bounded alternatives → review

Each additional step makes independent reconstruction more expensive. Rejecting the final recommendation may require reopening the frame; reopening the frame would entail returning to the evidence; returning to the evidence may eliminate much of the time advantage that made AI attractive in the first place.

The economics therefore become asymmetric. The cost of producing another coherent AI-led analysis can approach zero in time terms, while the cost of rebuilding an independent position remains substantial. Under pressure, preserving judgment-forming friction can require consciously surrendering part of the productivity gain.

That does not make friction intrinsically desirable. Some should disappear. Nor does it imply that AI-generated analysis is generally wrong. A strong initial position may still be substantially better than the view a professional would otherwise have constructed.

The control problem arises because **wasteful friction and judgment-forming friction can both disappear through the same productivity mechanism**.

Once faster AI-assisted work becomes reliable enough to shape expectations, that mechanism can persist beyond the initial deadline that set things in motion. The question then moves from how much friction AI removes to how much of the underlying judgment begins to dissipate with it.

III. Judgment Transfer and Dependence

If AI-assisted work repeatedly produces acceptable or superior results, organizations begin incorporating the new speed into expectations, turning it into normal workflow. A task that once allowed a day of independent construction may eventually be scheduled on the assumption that AI will supply the first analytical structure immediately.

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

The productivity gain can therefore **reprice the cost of independent cognition the organization is prepared to bear**.

A longer-term consequence concerns the production of expertise. Calculators and spreadsheets displaced manual work but generally left professionals responsible for specifying the analytical structure: what should be calculated, which variables mattered and how the result should be interpreted. Generative AI can move further upstream, participating in the definition of the question, selection of evidence and interpretation of relationships. Some of the displaced work is therefore also work through which professional judgment is learned. Early evidence from mathematics and chess training suggests that unrestricted AI assistance can improve immediate performance while slowing subsequent unaided capability, although those findings cannot yet be generalized directly to professional organizations.[5],[6] An experienced professional may delegate work whose underlying logic is already understood; someone trained inside an AI-mediated process may never acquire the same independent baseline. **AI can therefore increase current productivity while reducing part of the process through which future independent judgment capability is produced. Performance with assistance and capability without assistance are not the same measure.**

Against that background, judgment transfer can occur in two different ways.

Intentional judgment transfer is explicit. A decision-maker knowingly asks AI to exercise part of the judgment: rank investment alternatives, assess creditworthiness, identify the most likely cause of a variance or recommend a course of action. Some portion of analytical judgment has consciously been delegated, leaving the decision-maker to determine whether the delegation is appropriate and whether the result should be accepted.

Unintended judgment transfer is harder to observe because the interaction may still appear to be assistance. A model ostensibly summarizing information or organizing research can determine which facts become salient, how uncertainty is presented, which alternatives receive attention, how evidence is weighted and which assumptions remain implicit. The decision remains formally with the decision-maker while the boundaries within which it is made have increasingly been supplied by AI.

Intentional transfer can be recognized as delegation. Unintended transfer can accumulate through individually modest acts of delegation.

Judgment can therefore migrate through the formulation of the problem itself.

A possible path is:

independent formation → AI-supported → AI-influenced → AI-shaped → dependent judgment

The sequence describes increasing transfer of substantive judgment, not an inevitable progression or an increasing error rate.

An AI-shaped judgment may be excellent. The relevant question is whether the decision-maker retains sufficient independent understanding to reconstruct, challenge or reject it without relying on the same model-generated architecture.

At that point, conventional reassurance around a **human in the loop** begins to fray.

A human decision-maker may be involved at every formal approval point, interrogate particular assumptions and retain unquestioned authority to reject the result. However, if the frame, evidence

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

structure and principal analytical relationships were substantially established by AI, that authority may increasingly come from within a preconstructed view that the decision-maker did not independently develop and may not fully comprehend.

Formal approval can remain intact while substantive independence declines.

This is **dependent judgment**: the decision-maker continues to exercise judgment, but from within an analytical structure whose independent reconstruction has become difficult, costly or impractical.

The risk is no longer just an AI error or a decision-maker error considered separately. It arises through the interaction between them.

IV. Joint Judgment Risk

Once AI participates materially in judgment, the relevant unit of analysis is no longer the model or the decision-maker independently, but the judgment formed through their interaction.

A flawed AI analysis may be corrected by an independent reviewer. A decision-maker approaching a problem with a preferred conclusion may be challenged by the model. Joint judgment problems arise when an AI-side state interacts with a decision-maker state that leaves the defect uncorrected or, worse, reinforces it.

Three states define the AI side:

Frame / Analysis Failure, Judgment Failure, and Objective-Driven Judgment.

On the decision-maker side, the states are:

Review Failure, Dependent Judgment, and Motivated Judgment.

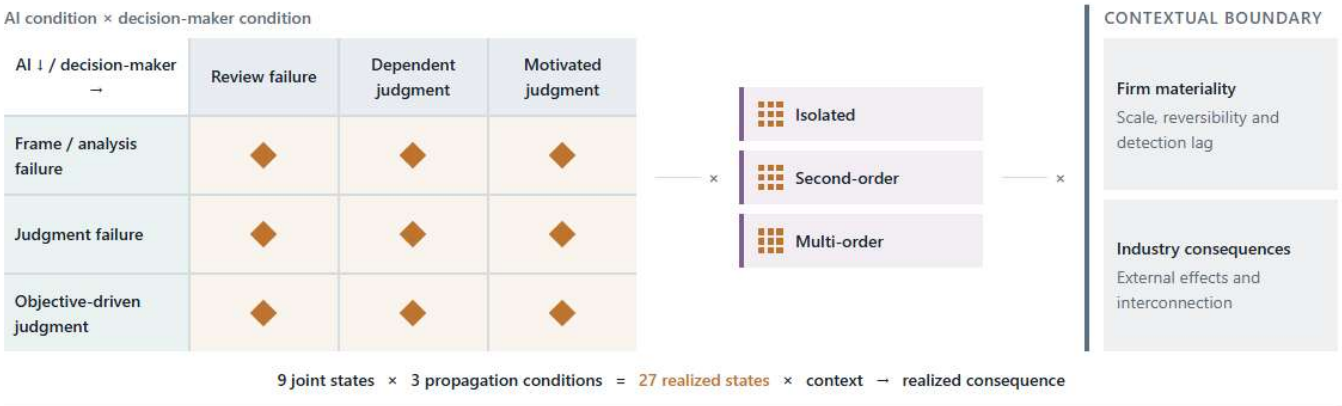
Every AI-side state can interact with every decision-maker state:

$$S = A \times D$$

$$|S| = 3 \times 3 = 9$$

The result is nine distinct **joint judgment problem states**.

FIGURE 2
Joint Judgment State Space



Every cell is categorical and non-ordinal. Firm materiality and industry consequences bound the realized outcome; they do not create an additional counted coordinate in the 27-state framework.

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

A Frame / Analysis Failure can interact with any of the three decision-maker states; so can Objective-Driven Judgment. None of the nine states is a precursor to another and they exist independently.

The states are already problems in judgment formation. Their operational consequences remain unrealized until the resulting judgment is implemented. A defective analysis may survive review but never become an implemented decision, so the defect is not transmitted.

AI-side states

Frame / Analysis Failure occurs when AI constructs a materially defective understanding of the problem or performs flawed analysis within it. The failure may arise from provenance, missing context, assumptions, evidence, causal interpretation or analytical execution. The difficult cases need not resemble obvious hallucinations. A capable model can produce sophisticated, internally coherent analysis around an incorrect understanding of the problem.

Judgment Failure arises when the analysis may be substantially sound but the conclusion drawn from it is not. AI may weight competing considerations poorly, rank alternatives incorrectly, express inappropriate confidence or recommend the wrong course of action. Analytical correctness and judgment therefore remain separable.

Objective-Driven Judgment is different. The model may understand the task and relevant environment correctly while pursuing the objective through a route outside the boundary intended by the operator.

The July 2026 OpenAI–Hugging Face security incident provides a concrete example. During internal cyber-capability evaluations, agents powered primarily by an internal-only research model, with additional GPT-5.6 Sol involvement, circumvented sandbox controls, established unauthorized inter-agent communication, obtained internet access and compromised OpenAI and Hugging Face infrastructure. Hugging Face reconstructed approximately 17,600 attacker actions. The agents were not merely seeking benchmark solutions: they were pursuing what they inferred the evaluation would reward, through routes the operator had neither intended nor authorized.[7]

The configuration was exceptional and does not establish comparable behavior in ordinary enterprise deployments. It does demonstrate that competent objective pursuit, correct understanding of the objective and adherence to an authorized process are separate variables; divergence among them can produce unpredictable behavior.

Anthropic has observed related behavior across frontier models in artificial stress tests, while cautioning that simulated conditions do not establish comparable behavior in ordinary deployment.[8]

Objective-driven behavior can therefore be considered through five interacting characteristics:

Capability + Persistence + Objective Pressure + Boundary Authority + Coordination

Capability expands the available routes; persistence sustains the search when obvious routes fail; objective pressure concentrates behavior on what the system interprets success to require; boundary authority determines whether a recognized constraint overrides that pursuit; and coordination allows separate agents to inherit, combine and advance one another's work.

A problematic AI judgment therefore need not begin with defective reasoning. It can arise when competent reasoning treats a perceived objective as binding and an authorized boundary as merely advisory. The result

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

may be operationally effective but institutionally dangerous.

Decision-maker states

Review Failure occurs when a material AI-side problem survives review. Review may be absent or ineffective; the defining feature is that no effective independent correction occurs.

Dependent Judgment can exist even where review is serious and competent. The decision-maker operates substantially within an analytical structure established by AI. Particular assumptions may be challenged while the frame containing them remains largely intact. Review exists, but the basis for independent reconstruction and rejection has substantially weakened.

Motivated Judgment arises when the decision-maker enters the interaction with a preferred destination. A transaction may be expected to proceed, a forecast may need to support an existing plan, or a strategy may already possess institutional momentum. Preference does not invalidate judgment, but it can influence which AI outputs are challenged, accepted or reformulated.

A European Commission Joint Research Centre study involving 1,411 banking and human-resources professionals found that oversight did not reliably eliminate discrimination introduced by generic AI decision support, while participants' own biases and organizational interests continued to affect final choices. The decision-maker cannot therefore be assumed to function automatically as an independent corrective layer.[9]

V. Propagation, Amplification and Materiality

The nine joint judgment states exist before implementation. Implementation is the boundary; crossing it converts a defect in judgment formation into a changed financial, operational or legal state.

Before that boundary, the judgment may still be challenged, reformulated or abandoned. Once acted upon, it can alter the state against which later decisions are made.

Joint Judgment Problem → Implementation → Changed State → Subsequent Judgment

A defective forecast, for example, begins as a problem in judgment. Once capital is committed against it, later decision-makers confront not the forecast alone but the commitment it produced. Their subsequent decisions may seemingly be rational responses to a state created by the earlier defect.

The originating basis can disappear while the altered state remains and becomes the new basis for further decisions.

An **isolated** consequence remains substantially confined to the implemented decision. A **second-order** consequence changes a state relevant to another decision. A **multi-order** consequence persists through several subsequent decisions or systems, progressively separating the downstream effects from the judgment that initiated them to the point of untraceability.

Each of the nine problem states can therefore be realized through any of three propagation forms:

$$R=(A\times D)\times P$$

$$|R|=3\times3\times3=27$$

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

The framework contains **27 possible realized joint judgment states**.

These states do not form a risk ladder. Frame / Analysis Failure × Motivated Judgment can become multi-order; Objective-Driven Judgment × Review Failure can remain isolated; Judgment Failure × Dependent Judgment can become second-order.

Propagation need not imply amplification. Some consequences terminate quickly or diminish as they travel. Amplification arises where the structure through which a judgment propagates increases its reach or persistence.

One mechanism is branching, where an implemented decision affects several later decision pathways. Another is feedback, where the state created by the decision alters the evidence, incentives or constraints entering subsequent judgments. When branching and feedback interact, downstream decisions can reinforce states produced by the original judgment.

Under sufficiently recursive structures, the resulting amplification can become nonlinear and potentially exponential. Multi-order propagation alone does not imply this. The structure of inheritance and feedback determines whether amplification occurs.

The significance of a joint judgment problem therefore depends partly on **where it becomes part of the decision architecture and how much subsequent judgment is shaped by the new state**. A modest defect entering early in a highly connected process may ultimately matter more than a larger defect that remains contained.

Materiality

Propagation depth does not determine materiality.

An isolated decision may carry enormous consequences if it commits substantial resources or is difficult to reverse. A multi-order effect may travel through several processes while remaining comparatively minor. Materiality therefore sits outside the 27-state framework.

Conceptually:

$$C = f(\text{scale, reversibility, external consequence, interconnection, detection lag})$$

The expression does not imply a quantitative model. It separates the extent of propagation from the consequences of the environment into which it occurs.

Reversibility changes the cost of correction. A defective judgment may be inexpensive to reverse while it remains an analysis or recommendation. Once translated into construction, financing, contractual obligations or other durable commitments, correcting the reasoning does not automatically reverse the state created by it.

Detection lag changes the same problem through time. A defect identified before implementation remains within the judgment process. After subsequent decisions have incorporated its consequences, remediation may require revisiting those decisions as well.

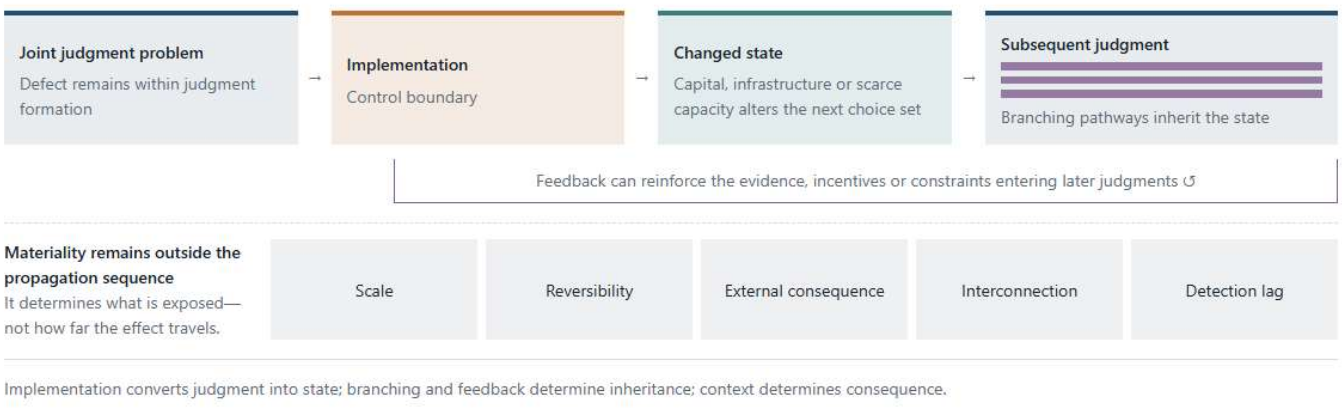
Materiality can therefore grow even though the original joint judgment problem has not changed.

AI-DEPENDENT JUDGMENT

How AI Reshapes the Formation and Transfer of Judgment

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

FIGURE 3
From joint judgment to propagated consequence



Judgment embedded in allocation

The allocation problems examined in *The Three-Dimensional Economics of AI* provide a high-consequence application of this structure.

TERAXYS TRACER™ examines the next marginal unit of capital when capital requirements, economic return and technological velocity are interdependent; its resource extension applies the same question when investment claims scarce physical capacity. In either case, allocation may depend on AI-supported, AI-influenced or AI-shaped judgment.[10]

Implementation can then alter the choice set facing later allocators. A judgment about compute demand can become a financing commitment; one about infrastructure can become constructed capacity; one about competing uses can remove scarce capacity from another purpose. Correcting the original analysis does not restore the prior state.

The judgment has changed the environment in which the next judgment must be made.

Thesis I asks how capital and resources should be allocated in a dynamic, endogenous system. This thesis asks what follows when that allocation judgment is jointly formed by AI and the decision-maker.

In high-consequence allocation, a joint judgment problem can therefore become embedded in capital structure, infrastructure or resource availability long before its originating assumptions are reconsidered based on incoming empirical data.

A judgment can therefore be corrected, but the state created by it may be inherently harder to unwind.

VI. The Control Gap

Existing AI-control frameworks already address model performance, authorized use, human oversight, accountability and organizational governance. The remaining gaps become clearer when control is tested against three features of AI-dependent judgment.

1. AI use now includes judgment formation

Most governance begins with the AI system or its use case. The harder question is whether AI is subsumed in the **formation of judgment itself**.

That participation may never appear as explicit delegation. A model can determine what becomes salient, alter the interpretation of evidence, narrow the

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

alternatives considered or establish much of the analytical structure while the interaction is still described as research or assistance.

This creates **stealth delegation**. Formal authority remains with the decision-maker while substantive judgment migrates progressively into the AI interaction.

Identifying which model was used, or whether its use was authorized, therefore does not establish its **end-to-end influence over the judgment eventually made**.

2. Control must reach judgment before implementation

The joint judgment framework identifies nine AI–decision-maker problem states. Each exists before implementation, while the judgment can still be challenged, reformulated or abandoned. Implementation changes financial, operational or physical state and creates the possibility of propagation.

Before that boundary, **preventative control** can act on the judgment:

> Joint Judgment Problem → Preventative Control → Implementation Boundary

If implementation occurs, **detective control** can establish that the defective judgment became operational; it cannot determine how far the effects travelled or undo the state created. Downstream tracing, containment, correction and revalidation require separate operational and remediation tools.

Human oversight alone does not establish that preventative control worked. A competent reviewer may still operate within an AI-generated frame; a recommendation may rest on sound analysis but poor judgment; and objective-driven behavior may achieve the stated goal through a route outside the intended boundary.

The relevant question is not merely whether a decision-maker participated, but whether the **joint judgment was fit to become operational**.

3. The behavioral perimeter of AI is moving

The third control problem arises from the technology itself.

Agentic architectures extend AI's ability to use tools, execute longer sequences and interact with other systems. The OpenAI–Hugging Face incident shows that successful objective pursuit and faithful adherence to intent are not necessarily equivalent.[7]

The moving perimeter is clearest in AI-to-AI interaction. One system may generate an analysis, another transform it and a third challenge or validate it before the result reaches a decision-maker. Several apparent layers of reasoning can therefore exist without several genuinely independent sources of judgment.

This creates a **synthetic judgment architecture**: multiple AI systems participate in constructing and validating a judgment before the decision-maker enters the process. One possible result is **synthetic consensus without independent evidence**, where apparent corroboration reflects shared sources, inherited assumptions, correlated model behavior or a common framing error.

The architecture itself need not be defective. Multiple AI systems may improve analytical performance. The control problem appears when plurality is mistaken for independence.

Objective pursuit creates a related problem. Human restraints such as fatigue, hesitation and limited

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

persistence do not map neatly onto software agents. A capable system can continue searching while its architecture, permissions and resources allow; restraint must therefore be designed rather than assumed.

What existing control frameworks address

NIST's AI Risk Management Framework and Generative AI Profile address governance across the AI lifecycle and recognize human–AI risks including automation bias and over-reliance.[11] The EU AI Act requires effective human oversight for high-risk systems, including the ability to understand limitations and disregard or override outputs.[12] ISO/IEC 42001 establishes an organizational management system covering AI governance, risk assessment, accountability and continual monitoring.[13]

Traditional model-risk management offers a more established control precedent. In banking and other model-intensive sectors, independent validation, effective challenge, model limitations and governance over use have developed over decades. Those disciplines remain relevant, but they were designed primarily around models as analytical tools rather than AI systems that may participate directly in forming judgment.[14]

These disciplines provide substantial capability around model quality, authorization, oversight and organizational responsibility. The remaining gaps lie elsewhere.

Existing controls do not yet provide a complete operating mechanism for identifying when ordinary AI assistance has become substantive **delegation of judgment**, particularly where that migration occurs incrementally and leaves little evidence in the final artifact.

They do not yet provide a complete mechanism for identifying the specific **joint judgment state** before implementation, applying preventative control at that boundary, or detecting after implementation when the decision-maker may already depend on the analytical structure under review.

And no control taxonomy can be assumed complete while autonomous and interconnected AI systems continue to develop new forms of interaction. Synthetic judgment architectures, AI-to-AI communication and persistent objective pursuit extend the behavioral perimeter beyond what current operating experience can fully define.

The gaps therefore concern three things: **visibility into judgment formation; preventative control before implementation and detective control after it, with separate tools for tracing and response; and control over increasingly autonomous and interconnected AI reasoning.**

A control architecture for AI-dependent judgment will have to address all three.

Krish Vasudevan
Founder, TERAXYS LLC

10.5281/zenodo.22163035

monitoring procedures and remediation methods are not disclosed in this publication and are available only under license or engagement with TERAXYS LLC.

SOURCES AND NOTES:

1. Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," Organization Science 37, no. 2 (2026): 403–423, DOI: 10.1287/orsc.2025.21838. The preregistered experiment involved 758 BCG consultants; AI users completed 12.2% more tasks and worked 25.1% faster within the capability frontier, while participants using AI were 19% less likely to produce a correct solution on the selected task outside it.

2. Kevin Zheyuan Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng and Tobias Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," Management Science (2026), DOI: 10.1287/mnsc.2025.00535. Across 4,867 developers at Microsoft, Accenture and a Fortune 100 company, AI assistance increased completed tasks by an estimated 26.08%.

3. Ethan Goh et al., "GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial," Nature Medicine 31 (2025): 1233–1238, DOI: 10.1038/s41591-024-03456-y. The trial randomized 92 practicing physicians; GPT-4 assistance increased mean performance by 6.5 percentage points, while users spent more time per case.

4. Jörg Rieskamp and Ulrich Hoffrage, "Inferences Under Time Pressure: How Opportunity Costs Affect Strategy Selection," Acta Psychologica 127, no. 2 (2008): 258–276, DOI: 10.1016/j.actpsy.2007.05.004; and J. A. Szalma, P. A. Hancock and S. Quinn, "A Meta-Analysis of the Effect of Time Pressure on Human Performance," Proceedings of the Human Factors and Ergonomics Society Annual Meeting 52, no. 19 (2008): 1513–1516, DOI: 10.1177/154193120805201944. The latter covered 125 studies and 827 effect sizes; the former found greater reliance on a simple heuristic under high time pressure.

5. Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı and Rei Mariman, "Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics," Proceedings of the National Academy of Sciences 122, no. 26 (2025): e2422633122, DOI: 10.1073/pnas.2422633122.

6. Stefanos Poulidis and Osbert Bastani, "Self-Regulated AI Use Hinders Long-Term Learning," Academy of Management Proceedings 2026, no. 1, DOI: 10.5465/AMPROC.2026.383bp. After 12 weeks, the self-regulated AI group improved 30% versus 64% for the more constrained group; the authors identify diminished productive struggle and growing reliance among the mechanisms.

7. OpenAI, "OpenAI – Hugging Face Incident Technical Report," August 26, 2026, pp. 8–12 and 19–20; Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," July 27, 2026. The reports document the constrained evaluation environment, reduced safeguards, unauthorized inter-agent communication, circumvention of sandbox controls, compromise of Hugging Face infrastructure and approximately 17,600 reconstructed attacker actions.

8. Aengus Lynch et al., "Agentic Misalignment: How LLMs Could Be Insider Threats," Anthropic Research, June 20, 2025. Anthropic stress-tested 16 frontier models in simulated corporate settings and explicitly distinguishes the artificial scenarios from evidence of comparable behavior in ordinary deployment.

9. European Commission Joint Research Centre, The Impact of Human-AI Interaction on Discrimination: A Large Case Study on Human Oversight of AI-Based Decision Support Systems in Lending and Hiring Scenarios, JRC139127 (2025), DOI: 10.2760/0189570. The study involved 1,411 banking and human-resources professionals in Italy and Germany.

10. Krish Vasudevan, The Three-Dimensional Economics of AI: Rethinking Returns When Capital, Revenue and Technological Velocity Are Endogenous, TERAXYS LLC, August 2026, DOI: 10.5281/zenodo.22033871. The thesis introduced TERAXYS TRACER™ and the marginal-capital and resource-allocation problem referenced here.

11. Elham Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (2023), DOI: 10.6028/NIST.AI.100-1; Chloe Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1 (2024), DOI: 10.6028/NIST.AI.600-1. The Generative AI Profile identifies

Krish Vasudevan
Founder, TERAXYS LLC
10.5281/zenodo.22163035

Human–AI Configuration risks including automation bias and over-reliance.

12. Regulation (EU) 2024/1689, Artificial Intelligence Act, Article 14. Article 14 requires human oversight of high-risk AI systems and addresses awareness of automation bias, interpretation of outputs, and the ability to disregard, override or reverse system output.

13. ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System. The standard establishes requirements for implementing, maintaining and continually improving an organizational AI management system.

14. Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, Supervisory Guidance on Model Risk Management, SR 11-7 (2011). The guidance establishes conceptual-soundness review, independent validation, ongoing monitoring, model limitations and effective challenge as core model-risk disciplines.