THE ARCHITECTURE OF AI CONTROL

Governing the Point of Consequence

KRISH VASUDEVAN
Founder, TERAXYS LLC

TERAXYS LLC · SEPTEMBER 2026

KRISH VASUDEVAN
10.5281/zenodo.22253085

I. The Control Mismatch

AI is increasingly changing where consequential judgment is formed. A model can now exercise preliminary judgment before a decision-maker sees either the output or the causal chain that produced it. It can shape the frame through which the problem is understood, select the sources considered and structure the basis on which facts, assumptions and alternatives are assessed. The decision-maker then exercises a further layer of judgment over an intellectual structure that may already have been materially formed by AI. Error can therefore arise within either layer, pass from one into the next, or be reinforced through subsequent judgment without becoming sufficiently visible. Formal approval may identify the human decision-maker, but it neither reconstructs the successive formation of judgment nor establishes where an error of fact or judgment first entered it. The control object must consequently extend beyond the system, the output and the act of review to the formation of judgment across AI and decision-maker layers, and to the detection of error within that structure.

Existing assurance objects

Existing control frameworks provide substantial assurance over the objects they were designed to govern. Traditional IT controls establish whether systems and processes operate within defined boundaries; model-risk disciplines reach further into assumptions, conceptual soundness and effective challenge; contemporary AI frameworks extend oversight to the development, deployment and human use of AI.[5,6] Some also address automation bias and the capacity of a decision-maker to interpret or reject an AI output. Yet these controls largely examine the model and the decision-maker as separate objects. Where AI has exercised preliminary judgment and the decision-maker subsequently judges within the intellectual structure it created, assurance over either layer in isolation does not reconstruct the synthesis through which judgment was formed. Nor does it establish whether an error originating in one layer was detected, inherited or reinforced by the next.

Legal force and operational reach

The NIST AI Risk Management Framework is expressly voluntary, and organizations may adopt its practices selectively; its publication therefore neither assures that its practices operate within any particular organization nor establishes the extent to which they have been implemented.[1] The EU AI Act occupies a different position: it is law, and significant provisions became enforceable on 2 August 2026. Yet its application remains phased. The requirements governing many high-risk systems—including the human-oversight provisions most relevant to consequential judgment—will apply from 2 December 2027 for Annex III systems and 2 August 2028 for AI embedded in regulated products.[3,4] That implementation schedule lags the rate at which AI systems and their institutional uses are changing. Legal authority and operational reach are therefore separate questions: a voluntary framework or a legally binding regime matters only to the extent that the relevant controls are actually implemented. The existence of a control framework does not, by itself, establish the degree of control achieved. In a high-complexity environment, the difference between a framework's existence and its operational implementation is itself a material control gap.

Coverage and technological pace

The more consequential limitation is one of coverage. NIST addresses human–AI interaction, monitoring and the risks of over-reliance, while the EU AI Act requires human overseers of high-risk systems to understand their limitations, interpret their outputs and remain capable of disregarding or overriding them.[23] Revised U.S. model-risk guidance preserves conceptual-soundness review and effective challenge, but expressly excludes generative and agentic AI models from its scope.[6] These are meaningful protections, but they principally govern systems, models and the interaction between an AI system and the person reviewing its output. They do not make the successive formation of judgment itself the control object: how preliminary judgment by AI established the frame and evidentiary basis, whether the decision-maker exercised independent judgment within that structure, or whether an error

10.5281/zenodo.22253085

passed between the two undetected or was reinforced. The same problem reaches an unprecedented level of complexity as judgment is distributed across multiple AI systems, persistent processes or objective-driven architectures before reaching a decision-maker at all. Control development must therefore contend not only with incomplete coverage but with a moving control perimeter. NIST is already revising AI RMF 1.0, while key EU human-oversight requirements remain ahead in the implementation schedule, even as the architectures through which AI participates in judgment continue to evolve. The question is no longer simply whether AI is governed, but whether control reaches the formation of consequential judgment involving AI with sufficient speed and depth.

II. Judgment Moves Upstream

The control mismatch described above has emerged as AI has moved from supporting execution and analysis to shaping the intellectual basis of decisions. Formal decision authority may remain unchanged at visible control points even as substantive judgment migrates upstream. Without adequate data, organizations cannot reliably estimate the cumulative effects of judgment formed at unobserved or unsupervised points in the process. The transfer occurs through the intellectual work that precedes the decision—framing the problem, determining relevance, interpreting evidence and structuring alternatives—so that the decision-maker may retain authority over a judgment whose foundations were substantially formed by AI. As developed in Thesis II,[7] declining friction can deepen this transfer through dependence, stealth delegation, objective-driven judgment and increasingly synthetic judgment architectures. Substantive judgment can therefore migrate before the organization recognizes that the locus of decision formation has changed.

III. The Judgment-Control Boundary

AI participation as a control event

Once AI contributes substantively to consequential judgment, control can no longer be satisfied by merely verifying that the decision-review process occurred. It must establish how substantive judgment was distributed within that process and whether that distribution remained within the authority assigned to the decision-maker and the permitted role of AI.

An output may arise from very different underlying structures: AI may supply discrete analysis within a frame established by the decision-maker, materially influence the frame and evidentiary architecture within which judgment is exercised, or shape the judgment so extensively that the decision-maker is principally evaluating a conclusion already constructed for them. These distinctions alter both the assurance that review can provide and the assurance the process requires. Current frameworks do not reliably determine whether independent judgment was preserved rather than reduced to affirmation of an AI-formed position. Control must therefore establish where AI participated, what part of the judgment it formed, how deeply that contribution shaped subsequent reasoning, the prospective exposure associated with that participation, and whether it was intended and authorized. The categories AI-supported, AI-influenced and AI-shaped serve this purpose as an observability taxonomy: they describe the scope of participation, not the severity of the resulting risk.

Transfer and retained independence

That inquiry cannot be confined to a single decision. Repeated AI participation can progressively transfer substantive judgment from the decision-maker to the systems on which they rely. As AI increasingly supplies interpretation and alternatives, the decision-maker may remain capable of reviewing an output while becoming less capable of reconstructing the analysis or reformulating the judgment independently. Transfer is therefore measured not only by what AI contributed to a particular decision, but by what the decision-maker can still do without AI decision support. A series of apparently well-reviewed decisions may cumulatively narrow independent judgment as AI-formed frameworks become the starting point for subsequent analysis. Control must therefore assess both the increasing share of judgment supplied by AI and the

KRISH VASUDEVAN
10.5281/zenodo.22253085

decision-maker's retained capacity for independent reconstruction, challenge and reformulation.

Nine joint-judgment states

The nine joint-judgment states introduced in Thesis II define the second control problem. Frame / Analysis Failure, Judgment Failure and Objective-Driven Judgment can each interact with Review Failure, Dependent Judgment or Motivated Judgment, producing nine distinct structures of defective joint judgment. Each state exists before implementation, while the judgment remains available for challenge, reconstruction, reformulation or abandonment. The control requirement is to identify the state present, establish how the AI and decision-maker contributed to producing it and determine the preventative response appropriate to the defects in that judgment.

This locates the defect in judgment formation before implementation and preserves the opportunity for preventative action.

The implementation boundary

The implementation boundary separates control over judgment from control over its effects. Before implementation, the nine joint-judgment states remain problems in judgment formation, and preventative control can act directly on the judgment by requiring challenge, reconstruction or reformulation, restricting AI participation or preventing the judgment from becoming action. Implementation changes the financial, operational or physical state of the organization and creates the possibility of propagation. Once that boundary has been crossed, detective control can establish that the defective judgment became operational, but it cannot determine how far the effects travelled, contain them or undo the state created. Downstream tracing, containment of effects, correction and revalidation require separate operational and remediation tools. Control must therefore distinguish the point at which defective judgment can still be prevented from becoming consequence from the point at which consequence must be investigated and addressed.

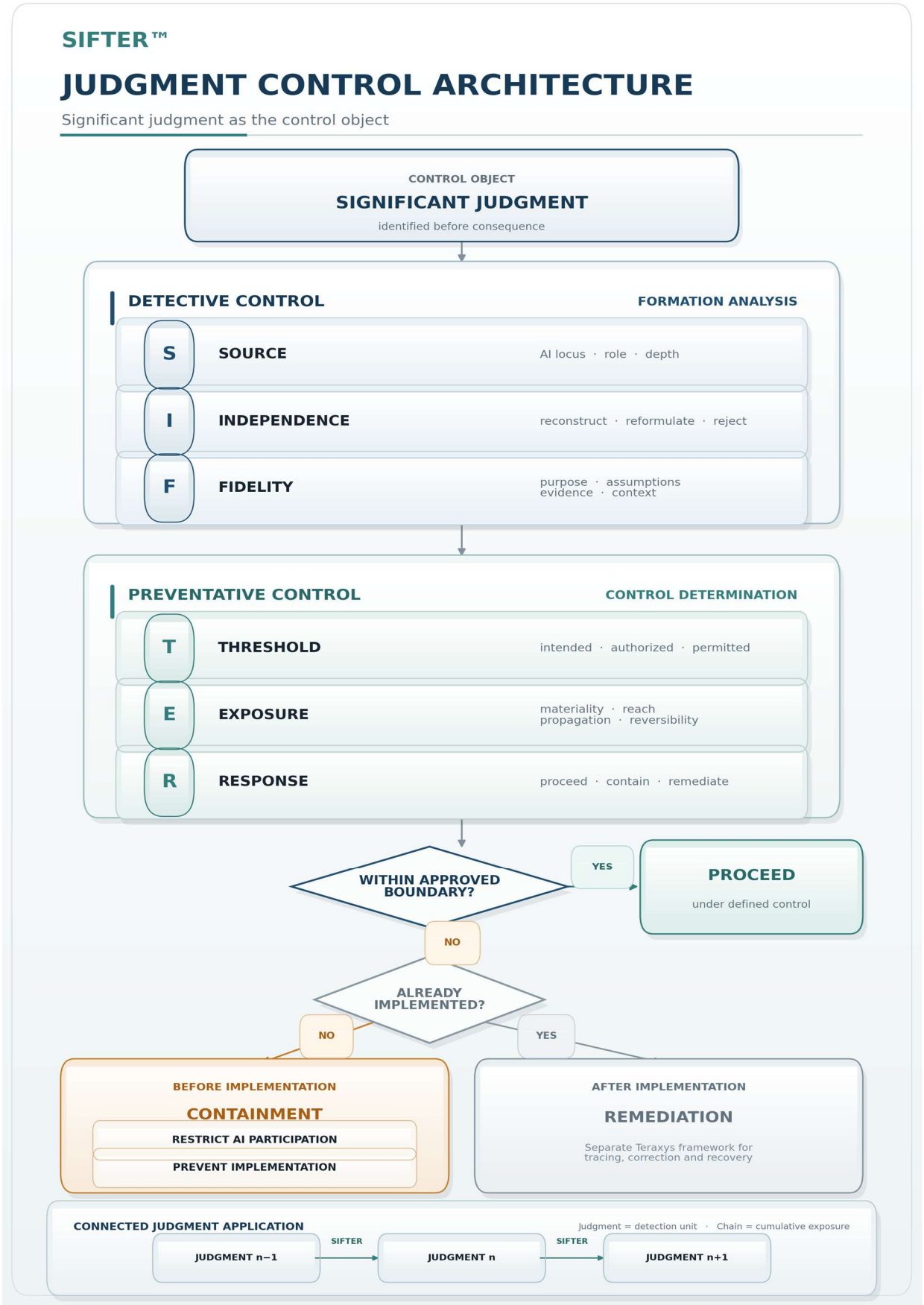
IV. SIFTER™: Judgment-Specific Control

The implementation boundary defines when the necessary control response changes, but it cannot define where control begins or what form it should take. A judgment-specific control layer must first identify every significant judgment within an organizational process and determine how that judgment is being formed: by a decision-maker using AI as support, through material AI influence over the frame or analysis, or within a structure substantially shaped by AI. This is the detective-control obligation. It reveals the judgment configuration before an error, breach or consequence has occurred. Once detected, the organization can determine whether AI participation was intended and authorized, whether the decision-maker retains the capacity to reconstruct and independently reformulate the analysis, and whether the judgment remains within the approved control boundary. Preventative control follows from that assessment. Where the detected configuration is inconsistent with the authorized form of judgment, control must act before implementation by requiring independent reformulation, restricting the AI contribution or preventing the judgment from becoming action. Detection and prevention therefore address different questions: the first establishes what form of judgment exists; the second determines whether that form may proceed. Neither can be reduced to a single model test or human-in-the-loop requirement. Together they require a control architecture capable of examining judgment formation across several distinct but connected dimensions.

Teraxys organizes this architecture through SIFTER™, a six-dimensional control framework applied to each significant judgment identified within an organizational process. It moves from detecting how judgment has been formed to determining whether that judgment may proceed and what intervention is required.

KRISH VASUDEVAN
10.5281/zenodo.22253085

FIGURE 1 · **SIFTER™ JUDGMENT CONTROL ARCHITECTURE**



10.5281/zenodo.22253085

Source

Source identifies the locus and degree of AI participation in the judgment under examination. It distinguishes whether AI supported, influenced or shaped the judgment.

Independence

Independence determines whether the decision-maker can reconstruct the analysis, reformulate it without the intellectual structure supplied by AI and reject its conclusion on an independently formed basis.

Fidelity

Fidelity evaluates whether the original purpose, assumptions, evidence, limitations and relevant institutional context remain intact as judgment moves between AI and human layers.

Threshold

Threshold establishes whether the detected configuration remains intended, authorized and permitted to proceed. It defines the point at which preventative control must restrict AI participation or prevent implementation.

Exposure

Exposure assesses the materiality, reach, propagation and reversibility of the consequences associated with implementation. It establishes the prospective organizational impact against which the required control response is determined.

Response

Response determines whether the judgment may proceed under the existing control environment or requires intervention. Where intervention is required, SIFTER separates containment from remediation according to whether implementation has occurred.

Containment

Containment operates before implementation and prevents an unauthorized form of judgment from becoming action. It can take one of two principal forms: restriction of AI participation or prevention of implementation.

Restriction of AI participation

Restriction reduces AI participation to an authorized form—for example, by removing an AI-formed frame, limiting AI to discrete analytical support or requiring the decision-maker to reformulate the judgment independently. The revised configuration must then be reassessed before it proceeds.

Prevention of implementation

Prevention blocks the judgment from becoming action where restriction cannot restore an acceptable judgment structure or where the exposure exceeds the approved boundary. The underlying business action may still proceed through a newly constituted and authorized judgment process.

Remediation

Remediation begins after implementation, once the judgment has produced consequences. SIFTER identifies the remediation branch and defines the exposure to be addressed; a separate forthcoming Teraxys framework will establish the methods for judgment reconstruction, causal tracing, containment of effects, correction and recovery.

Together, the six dimensions establish how significant judgment was formed, test that form against the approved control boundary, assess the exposure associated with implementation and determine the required disposition: proceed, contain before implementation or enter remediation after implementation.

Cumulative application across judgment chains

The individual significant judgment remains the unit of detection, while the connected judgment chain becomes the unit of cumulative exposure. The nine joint-judgment states characterize the AI–decision-maker configuration at each point in that chain; propagation order and materiality determine how those configurations interact as outputs, assumptions and consequences are inherited by later judgments. SIFTER addresses this accumulation through the same six dimensions. Source, Independence and Fidelity identify

10.5281/zenodo.22253085

inherited frames, analyses or objectives and establish whether the next decision-maker has independently reconstructed them. Threshold and Exposure assess the current judgment together with its accumulated dependencies, including propagation depth, scale, interconnection, reversibility, external consequences and detection lag.

Where that cumulative configuration exceeds the authorized boundary, restriction requires independent reformulation before the judgment can serve as an input to the next decision, while prevention blocks implementation and the creation of a further order of effect. Where implementation has already occurred, the accumulated exposure enters the separate remediation branch. SIFTER thereby interrupts the mechanism of cumulation at the point of transition from inherited judgment to a new order of organizational effect.

Control before consequence

The control objective is to align formal authority with substantive command over the judgment exercised under that authority. Existing IT, model-risk and AI-governance disciplines provide assurance over systems, models and their use; judgment-specific control extends that assurance to the formation of the judgments on which the institution acts. Together, they preserve the practical capacity to understand, reconstruct and redirect the reasoning that produces organizational action as AI participation deepens. AI can therefore expand analytical capacity within terms that remain explicit, observable and institutionally governed. Accordingly, control in an AI-dependent institution begins before consequence: at the point where significant judgment is formed and authorized for implementation.

10.5281/zenodo.22253085

SOURCES AND NOTES:

1. *Elham Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (2023), DOI: 10.6028/NIST.AI.100-1. NIST describes the framework as intended for voluntary use and states that AI RMF 1.0 is being revised.*
2. *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; NIST AI RMF Playbook, MAP 2.2 and MEASURE resources addressing human oversight, human–AI configurations, cognitive bias and upstream system dependencies.*
3. *Regulation (EU) 2024/1689 of the European Parliament and of the Council, Artificial Intelligence Act, consolidated version of 27 July 2026, as amended by Regulation (EU) 2026/1744. Article 14 addresses human oversight of high-risk AI systems; the amended application schedule appears in Article 113 and related provisions.*
4. *European Commission, 'AI Omnibus enters into force,' 27 July 2026. The Commission states that rules for high-risk AI systems in Annex III apply from 2 December 2027 and rules for high-risk AI embedded in regulated products apply from 2 August 2028.*
5. *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.*
6. *Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation and Office of the Comptroller of the Currency, Revised Guidance on Model Risk Management, SR 26-2 (17 April 2026). The guidance supersedes SR 11-7, preserves conceptual-soundness review and effective challenge as core model-risk disciplines, and expressly excludes generative and agentic AI models from its scope.*
7. *Krish Vasudevan, AI-Dependent Judgment: Can Today's Control Frameworks Support AI-Dependent Judgment?, TERAXYS LLC, August 2026. Thesis II introduces the AI-supported / AI-influenced / AI-shaped taxonomy, the nine joint-judgment states and the propagation and materiality structure used in this thesis.*