Article

Not peer-reviewed version

The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains

Hiroki Naito *

Posted Date: 19 March 2026

doi: 10.20944/preprints202603.1507.v1

Keywords: human-in-the-loop; AI governance; moral crumple zone

 doi:10.20944/preprints202603.1507.v1

Article

The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains

Hiroki Naito

UTIE Research Institute (UTIE Instruments Inc.), Japan; h.naito@utie-instruments.com

Abstract

Current AI governance rests on an implicit assumption: that human-in-the-loop oversight can scale safely alongside generative AI systems. We argue that this assumption is structurally untenable. Three constraints hold simultaneously: legal liability remains with humans, human cognitive throughput has a biological ceiling, and economic pressures drive AI output velocity beyond that ceiling. When output velocity exceeds human processing limits, oversight becomes nominal and humans are reduced to what Elish (2019) termed "moral crumple zones". Unlike physical automation, where anomaly criteria are externally defined, generative AI requires supervisors to evaluate cognitive products (text, reasoning, analysis) against internally held standards. Through predictive-error minimization, repeated exposure to AI output patterns recalibrates these internal standards, degrading anomaly detection even when supervisors remain attentive. This degradation renders error detection structurally deficient, causing the penalty function that should restrain output expansion to remain dormant. Risks therefore accumulate invisibly and manifest as threshold shocks rather than gradual corrections. Under these dynamics, expected loss diverges with increasing output velocity, and the irreducibility of error probability in probabilistic systems ensures that model capability improvements cannot offset this divergence. We derive that rate-limiting AI output to within human processing capacity is the variable available for bounding expected loss, and propose a flow-design governance paradigm as a principled alternative to supervision-enhancing approaches. Specifically, we outline hard caps on daily case loads, batch-approval prohibition, mandatory friction in approval interfaces, and adoption-rate ceilings. The theoretical consequence is counterintuitive: as generative AI capability grows, its autonomous use in high-loss domains will contract.

Keywords: human-in-the-loop; AI governance; moral crumple zone

1. Introduction

Contemporary AI governance rests on a single implicit premise. Namely, the premise that so long as a human who bears ultimate judgment and responsibility remains within the decision-making loop (Human-in-the-Loop), the system can scale safely, no matter how rapidly or extensively the outputs of generative AI systems grow. The human oversight obligations for high-risk AI systems under the EU AI Act, the notion of human involvement in the U.S. NIST AI Risk Management Framework, and the single line found in corporate deployment guidelines stating that "final verification must always be performed by a human" all rest on the assumption that human cognitive capacity will continue to maintain substantive supervisory function relative to the system's output velocity. We argue, however, that this premise is structurally untenable. This argument is derived from the simultaneous satisfaction of three constraint variables. First, legal and social liability remains attributed to humans ($R = 1$). No legal framework in any major jurisdiction grants AI systems legal personhood, contractual standing, or criminal liability. $R = 1$ is not a mathematical variable but an institutional precondition within our model. Rather than being incorporated as a function of other variables, it serves as the boundary condition, namely that liability attribution is fixed to humans,

 doi:10.20944/preprints202603.1507.v1

upon which the entirety of the following discussion proceeds. Second, human cognitive processing capacity has a biological ceiling (C_max). The volume of information that a human can scrutinize, detect errors in, and render decisions upon per unit of time is finite, and this constraint does not qualitatively change through training or assistive tools. Even if AI-based assistive tools (automated checks, etc.) are introduced, the burden of ultimately verifying the output of those tools themselves remains, and so the constraint persists. Third, economic pressures continuously expand AI output velocity and scope of application. What matters is not that this expansion follows a specific function, but that it scales persistently and irreversibly beyond C_max ($V \rightarrow \infty$). When these three variables simultaneously hold, a natural consequence arises for the system. At the point where output velocity V exceeds the human processing limit C_max, oversight through Human-in-the-Loop structurally becomes a hollow formality. That is, humans are transformed into what Elish (2019) called the "Moral Crumple Zone," nominal bearers of responsibility who absorb all impact. Arguments corroborating this structural concern have already been raised from multiple directions. Carnat (2024) argued that automation bias in generative LLMs cannot be resolved by the human oversight requirements of Article 14 of the EU AI Act, and pointed to the paradox that technically improving the hallucination problem itself accelerates human overreliance. Horowitz & Kahn (2024) demonstrated, in an experiment involving 9,000 participants across 9 countries, that trust in AI systems and self-confidence are major factors driving automation bias.

These problems are not new. Bainbridge (1983), in Ironies of Automation, pointed out that the more advanced automation becomes, the more the tasks remaining for human operators paradoxically become only those two that are least suited to human cognitive characteristics: "tedious monitoring of a system that normally functions perfectly" and "intervention under extreme conditions when the system encounters an unknown failure." Perrow's (1984) Normal Accident Theory demonstrated that in highly complex and tightly coupled systems, accidents are not anomalies but structural consequences. Beck's (1986) risk society thesis argued that modern technological development inherently and systematically generates risks that exceed existing institutional capacities for control. These studies have accumulated primarily with physical automation systems, notably nuclear power generation, aviation, and chemical plants, as their objects of inquiry. What is at issue today, however, is that the same structural problem is being reproduced at an unprecedented scale and speed in the very core of human cognitive and intellectual work: text generation, summarization, decision support, and code generation. In physical automation, the object that humans entrust their monitoring to is the operation of machines. In generative AI, humans monitor text, media, and code output by AI. In other words, what humans must monitor are "cognitive and inferential products" of the same kind as those they themselves ordinarily produce, and as a consequence, the very cognitive processes that form the foundation of their monitoring are themselves exposed to the risk of transformation.

Conventional governance theory has relied on the assumption that society reaches equilibrium through continuous feedback, incrementally adapting institutions in response to the materialization of technological risks. Just as the mass occurrence of automobile accidents gave rise to traffic regulations, and the analysis of aviation incidents prompted the mandatory adoption of flight recorders, systems are said to gradually stabilize through repeated cycles of problem occurrence and institutional response. However, hidden within this historical analogy is the major premise that technological change and institutional change proceed on the same time scale. The pace of automobile and aviation development was constrained by physical limitations such as steel refining, engine testing, and the construction of large-scale factories, and therefore progressed in a roughly linear fashion, managing to barely synchronize with the speed of society's institutional adaptation. The output velocity of generative AI expands nonlinearly at a speed fundamentally different from that of past physical technologies, operating within an information space unconstrained by physical limitations. Meanwhile, the revision of legal systems necessary to accommodate this, the redesign of liability attribution, and the formation of social consensus are all bound by the friction of physical

 doi:10.20944/preprints202603.1507.v1

society, including legislative processes, bureaucratic apparatus, and judicial institutions, and there exists an absolute ceiling on the speed of these changes.

This paper develops the above structure as follows. In Chapter 2, we analyze the bidirectional collapse that scaling brings about under the three fundamental constraints. This encompasses the degradation of supervisory cognitive capacity (microscopic limit) and the boundedness of institutional adaptation speed (macroscopic limit). In Chapter 3, we reject the assumption of gradual adaptation through continuous feedback, and describe the dynamics by which risks materialize discontinuously as threshold shocks. In Chapter 4, as consequences of the above, we derive that flow-rate limitation of AI output in high-loss domains constitutes the rational equilibrium for expected-loss minimization, and present a rational explanation for why capability improvement necessarily entails usage contraction. In Chapter 5, we specify the limitations and scope of this study, and in Chapter 6, we state our conclusions. It should be noted that the domain this study addresses is one where losses upon error occurrence are enormous and where formal or mechanical verification of correctness is inherently impossible, that is, the domain of "high loss and high uncertainty." Everyday low-risk uses, and domains where external criteria can be established through statistical verification, fall outside the scope of this paper. The arguments of this study are valid only under these delimited conditions.

2. Fundamental Constraints in Scaling and the Limits of Cognition and Institutions

2.1. Definitions

The model in this paper is grounded in the following three variables. Each is a constraint that is effectively fixed under current technological, legal, and social conditions.

Constraint 1: Attribution of Liability (R = 1)

For damages arising from the outputs of a generative AI system, the entity that bears ultimate legal liability is a human, or a legal person controlled by humans. As of now, no legal system in any major jurisdiction of any nation grants AI systems themselves legal personhood, standing as a contracting party, or criminal liability capacity. Given that the fundamental function of legal systems, namely imposing penalties on liable parties and thereby correcting behavior, presupposes the existence of humans who possess consciousness and volition, R = 1 is a constraint inherent in the constitutive principles of legal systems.

Constraint 2: Upper Bound of Human Cognitive Processing Capacity (C_max)

There exists a biologically determined upper limit on the volume of information that a human can scrutinize, detect errors in, and render judgments of correctness upon per unit of time. The finite nature of attentional resources (Kahneman, 1973), the capacity constraints of working memory (Cowan, 2001), and the temporal decay of sustained attention (Warm, Parasuraman & Matthews, 2008) are findings already well established in cognitive science. Moreover, this upper limit does not qualitatively change through training or the introduction of assistive tools. Individual differences and variations in expertise do exist, and these introduce variation in the value of C_max, but they do not alter the order of the variable.

Constraint 3: Nonlinear Expansion of Output Velocity ($V \to \infty$)

The output velocity and scope of application of generative AI systems can be strongly predicted to continue expanding beyond C_max, driven by economic pressures such as decreasing per-unit cost, increasing processing speed, and securing competitive advantage. This acceleration is not subject to the constraints of physical manufacturing processes. Because generative AI output is completed entirely within information space, physical bottlenecks such as steel refining or factory construction do not exist. Each time the inference cost of a model decreases, the range of deployable business domains and the volume of outputs expands nonlinearly. In other words, the expansion of V is structurally sanctioned not only as a possibility of technological progress but also as a rational course of action chosen by firms under market competition.

 doi:10.20944/preprints202603.1507.v1

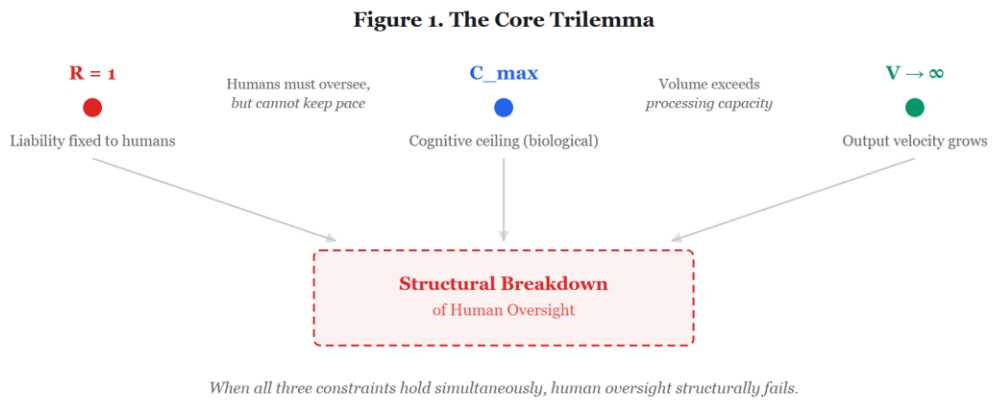


Figure 1. This figure illustrates the structural incompatibility of three simultaneously held constraints: the fixed attribution of human liability (R=1), the biological ceiling of human cognitive processing capacity (C_max), and the nonlinear expansion of AI output velocity ($V \to \infty$). When output volume persistently exceeds human processing limits, the human-in-the-loop safety mechanism structurally devolves into a hollow formality.

When these three constraints simultaneously hold, the following inequality comes to hold over time:

$$\mathbf{V(t) \gg C\_max, \quad with \quad R = 1}$$

In the subsequent sections of this chapter, we analyze the consequences implied by this inequality from two directions: the cognitive dimension (microscopic) and the institutional dimension (macroscopic).

2.2. Microscopic Limit

What the safety premise of Human-in-the-Loop implicitly assumes is that the cognitive capacity of supervisors is maintained at a constant level. That is, it assumes that the error detection rate of humans scrutinizing AI output remains constant even as the volume of outputs increases, or improves through the accumulation of experience. However, this assumption is beset by numerous problems. Bainbridge (1983) pointed out that as the level of automation rises, the role remaining for human operators paradoxically transforms into one that is increasingly difficult. In the context of generative AI, this dynamic takes on an even more severe form. Errors in physical automation manifest as deviations in physical quantities such as temperature or pressure. The range of normal values is objectively defined, and the deviation criteria for judging anomalies exist "extrinsically," external to the supervisor. By contrast, in the cognitive products (text and reasoning) that supervisors must inspect in generative AI, errors manifest as breakdowns in semantic or contextual coherence or as logical leaps. The deviation criteria for judging these as normal or anomalous depend "intrinsically" on the supervisor's own internal model. Because the criteria are intrinsic, a distinctive vulnerability arises here. The text and formats output by generative AI possess stronger statistical regularity than those produced by humans (this statistically stable output pattern is hereafter referred to in this paper as a template). Supervisors who are repeatedly exposed to AI output progressively learn and internalize this regularity as the "normal state," through the predictive-error minimization mechanism described by the Free Energy Principle (Friston, 2010) and Predictive Processing (Clark, 2013) models. As the brain adapts to the statistical regularity of AI and prediction error decreases, it is subjectively experienced at the cognitive level as heightened Processing Fluency (Alter & Oppenheimer, 2009), that is, a diminished sense of discomfort toward the output. As a result, the supervisor's standard for "normal output" (the intrinsic deviation criteria) can be redefined by AI templates. So long as the format and vocabulary conform to the AI's regularity, the brain becomes less likely to generate a sense of discomfort (prediction error signal) even toward subtle factual distortions or logical leaps contained within. Beyond the automation bias identified by Parasuraman & Riley (1997), what occurs here is a structural degradation of error detection ability through a

 doi:10.20944/preprints202603.1507.v1

transformation of the cognitive framework itself, even while the supervisor is paying attention. Supervisors, by becoming proficient in AI output patterns, may in fact improve their discrimination ability among the presented set of options (within-set discriminability in AISP; Naito, 2025). However, at the same time, the loss of a sense of discomfort exerts upward pressure on the hazard-miss rate, i.e., the rate of overlooking risk factors that exist outside the set of presented options. This dynamic is gradually being confirmed in recent large-scale reviews on automation bias. A study that systematically reviewed 35 studies from 2015 to 2025 reported that interventions such as explanation provision through XAI and trust-calibration feedback are ineffective at reducing automation bias (AI & Society, 2025). Empirical research in the medical domain has shown that non-experts are more vulnerable to automation bias in clinical decision support systems (Kücking et al., 2024). Furthermore, McCormick (2025) coined the term "interpretive debt" for the phenomenon in which the human judgment framework is progressively reorganized through sustained, high-coherence dialogue with AI systems, even when the AI system is functioning normally and not committing errors, and argued that this constitutes a risk pathway independent of conventional automation bias.

As a consequence, the effective cognitive processing capacity of the supervisor (C_eff) is not a static value that remains at the biological upper limit (C_max). Domain-specific expertise and independent judgment foundations formed prior to AI adoption can be reorganized through cumulative exposure to AI output. When this reorganization acts in the direction of aligning internal standards with AI modalities, it carries an inherent risk of weakening independent deviation detection capacity. What this means is that the effective error manifestation rate of the coupled system under the Human-in-the-Loop model is not a constant. An increase in output velocity (V) does not merely increase the number of error-occurrence trials N; it also structurally worsens the effective error manifestation rate of each individual trial by rewriting the supervisor's internal model. This constitutes a nonlinear risk amplification factor that is qualitatively different from the mere increase in the number of trials under a constant error rate assumed by conventional static risk models.

This transformation of the cognitive foundation does not stop at the degradation of error detection ability. It has already been empirically reported that lexical and syntactic diversity is significantly reduced in collaborative environments with LLMs (Padmakumar & He, 2023; Yakura et al., 2024). The reduction of vocabulary signifies the shrinkage of the conceptual tools through which the supervisor (human) segments the world and recognizes deviations. Whereas the degradation of supervisory capacity in traditional automation bias problems operates as a decline in attentional capacity, the degradation of supervisory capacity in generative AI operates as "an impoverishment of the cognitive vocabulary itself for recognizing deviations." This difference also follows directly from the distinction drawn in this paper between extrinsic and intrinsic deviation criteria.

The arguments of this study are not intended to constitute a completed longitudinal empirical study. What this paper presents is a structural conditional proposition, an argument of the form "if these constraints simultaneously hold, then the following consequences are derived." However, each constituent element of this study is already individually supported by existing empirical research. The accumulated research on automation bias in automated environments (Parasuraman & Manzey, 2010 onward), recent experimental findings demonstrating human overreliance on LLM output, and systematic evaluations of hallucination rates in generative AI all suggest that the potential degradation of C_eff and the persistence of $p > 0$ are empirically observed phenomena. We therefore do not comprehensively re-examine these empirical studies but instead present arguments that presuppose them.

2.3. Macroscopic Limit

Another line of defense regarding the sustainability of Human-in-the-Loop is the argument of institutional adaptation. The assumption that even if problems arise as technology advances, society will gradually adapt through legal reform, the development of industry standards, and the updating of organizational governance, eventually arriving at an equilibrium of safe system operation. This assumption has historical backing. The proliferation of automobiles gave rise to traffic laws and

 doi:10.20944/preprints202603.1507.v1

insurance systems, and the analysis of aviation accidents prompted the mandatory adoption of flight data recorders and international safety standards. However, the premise that the speed of technological development and the speed of institutional adaptation belong to the same time scale does not hold in the governance discussion of modern generative AI. We formally describe the relationship between the speed of technological development V(t) and the speed of institutional adaptation S(t). The output velocity of generative AI is not subject to physical constraints, but the speed of institutional change is entirely different. The process from bill drafting through deliberation, passage, and enforcement, or the process from the development to the implementation of industry standards, is constantly constrained by the friction of physical society: the deliberative capacity of legislative bodies, the processing speed of bureaucratic apparatus, the cost of consensus formation among stakeholders, and the establishment of judicial interpretation. Given that $\dot{V}$ increases in proportion to V while $\dot{S}$ has an absolute upper bound, under any initial condition, from some time point t* onward, V(t) − S(t) monotonically increases, and the gap diverges. In other words, institutions never catch up, and an "area beyond human supervisory reach" emerges between institutions and technology. Domains where AI output is exerting real-world impact yet the institutional frameworks governing it remain either undeveloped or already obsolete will expansively continue to exist.

2.4. The Relationship Between Supervisory Capacity and Accuracy

The microscopic and macroscopic limits are not independent. Delays in institutional adaptation impose an additional burden of ambiguity in judgment criteria on supervisors, which in turn can accelerate the decline of C_eff. Conversely, the degradation of supervisory capacity lowers the precision of problem detection and reporting, thereby weakening the informational foundation upon which institutions adapt. Because the two interact in this manner, they are not in a simple parallel relationship but may, under certain conditions, enter an amplifying relationship. In such cases, the gap between technological velocity V and institutional adaptation S does not merely widen; it manifests in the form of a delay in the visibility of problems themselves. As a result, states can arise in which risks are accumulating internally even during phases that appear stable on the surface.

3. Nonlinear Risk Accumulation and Discontinuous Feedback

Chapter 2 demonstrated, in straightforward logical terms, the structure in which human cognitive capacity and institutional adaptation speed are jointly unable to keep pace with the expansion of AI output velocity. This chapter discusses how risks accumulate under this structure and how they materialize.

3.1. The Assumptions and Limitations of the Continuous Adaptation Model

Many current AI governance theories implicitly presuppose the following feedback model. When errors occur in the operation of an AI system, they are detected, reported, and analyzed. As a result of this analysis, operational procedures are revised, or regulations are updated. These corrections are reflected in the system, thereby reducing the error rate, and the system progressively improves its safety. This continuous adaptation model has been widely adopted in quality control and safety engineering, and its effectiveness has been demonstrated across many industrial domains. This continuous adaptation model represents the temporal evolution of output velocity V as follows:

$$\mathbf{dV/dt = F(V) - g(E)}$$

Here, F(V) is the acceleration term driven by market competition, capital pressure, and technological optimization. Whether it is linear, strongly nonlinear, or involves self-reinforcing acceleration, the intensity itself is not what matters. The core premise of this model is the following point: namely, the premise that when errors occur, they are detected; cumulative error E is made visible; and the social penalty g(E) operates continuously. If negative feedback (regulatory tightening, litigation costs, reputational damage) increases smoothly, then even with a strong acceleration term F(V), ultimately

 doi:10.20944/preprints202603.1507.v1

F(V) = g(E)

will hold, and the system will reach an equilibrium point. That is, logistic-curve-like saturation will naturally arrive. This is the belief that existing governance theory has implicitly shared. Our argument is that the premises of this belief are structurally unsatisfied in the context of generative AI.

3.2. Invisibilization of Errors

The necessary condition for the continuous adaptation model to function is that when errors occur, they are detected in a timely manner and input into the feedback loop. However, the degradation of supervisory capacity (decline in C_eff) discussed in the previous chapter can introduce Systemic Bias into this detection process. Under supervisors who have adapted to AI output templates, subtle factual distortions and logical leaps tend to pass through the filter of attention and be processed as "normal output," so long as the format is consistent. If such bias in detection capacity persists, the following state may arise in the system's operational processes. If we denote the actually accumulated quantity of errors as E_actual and the quantity of errors detected and reported by supervisors as E_detected, a systematic divergence can arise between the two:

E_detected ≪ E_actual

Because the penalty function g operates only on errors that have been made visible within the organization, as long as visibility remains insufficient, the restraining effect is evaluated as smaller than the reality. This state engenders a deceptive sense of stability in organizational operations. Internal quality reviews and compliance checks are formally passed, and the error rate appears to be stable at low levels (KPIs appear to be improving), yet behind the scenes, a qualitative deterioration of judgment criteria is progressing. While no critical incidents surface, minor hallucinations and logical errors become routinized and accepted as standard system behavior. These realities are already being empirically confirmed in clinical domains. Asgari et al. (2025) reported a 1.47% hallucination rate and a 3.45% omission rate in LLM-generated clinical note summaries. While these are low percentages, under the conditions of routine high-volume processing, these errors accumulate. As an even more serious finding, Omar et al. (2025) reported that six major LLMs reproduced and expanded upon single pieces of false data (fabricated test values or fabricated signs) inserted into clinical vignettes at rates of up to 82%. In other words, unless supervisors independently verify the accuracy of input data, errors embedded in AI output templates are amplified as-is. This structure has now been confirmed.

3.3. Threshold Shocks

Accumulated errors do not remain invisible in perpetuity. Outputs that are treated as probabilistic errors within information space can, the moment they connect to physical space (the human body, infrastructure, institutional procedures), be converted into deterministic damages. Physical and legal damages generally produce qualitatively different consequences once a certain threshold is exceeded. For example, a minor error in a legal document prepared by generative AI does not manifest as an operational problem so long as it remains undiscovered. However, when the error is made visible through litigation or a regulatory investigation, its consequences extend beyond repair costs and materialize in the form of contract nullification, damages claims, and administrative sanctions. Accordingly, the magnitude of damage does not increase smoothly in proportion to the quantity of errors but leaps in stages upon the crossing of critical points. Furthermore, the materialization of a single incident can retroactively make visible errors of the same type that were latent within the same process. Through this cascading visibility, the penalty function g does not ramp up continuously but rather fires sharply once a certain critical error quantity E_critical is exceeded. The penalty function implicitly assumed by the continuous adaptation model presented in Section 3.1 is a continuous function in which g(E) rises smoothly with increasing E. By contrast, the dynamics our argument demonstrates take the following threshold structure:

g(E) ≈ 0 (E < E_critical) **g(E) → ∞ (E ≥ E_critical)**

 doi:10.20944/preprints202603.1507.v1

And this threshold-type activation irreversibly distorts the temporal structure of feedback. If corrections are introduced at the stage of small problems, as the continuous adaptation model envisions, the system can reach equilibrium without experiencing large-scale breakdown. However, under threshold dynamics, for the extended period leading up to E_critical, the penalty remains at a local and superficial level and does not provide an occasion for revisiting the design of the system as a whole. Corrections are confined to the handling of individual cases and do not connect to structural adjustment. As a result, the system cannot sufficiently activate a circuit for "learning at the stage of small problems." Consequently, rather than operating as gradual improvement, feedback carries a pressure that tilts the system toward firing as an abrupt institutional reaction after the critical point is exceeded. The intensity of this pressure depends on the domain and institutional conditions, but the concern of this paper lies not in its quantitative evaluation but in the identification of structural conditions capable of generating discontinuous feedback.

Figure 2. Error Accumulation and Threshold Shock

Figure 4. This figure demonstrates how the degradation of effective cognitive capacity creates a widening "invisible error gap" between actual accumulated errors (E_actual) and detected errors (E_detected). Consequently, the institutional penalty function g(E) fails to operate continuously. Instead, it remains dormant during the accumulation phase and fires only as a reactive, discontinuous shock once a critical threshold (E_critical) is breached in physical space.

3.4. AI Governance Failure as a Normal Accident

Integrating the dynamics described above, the following scenario is derived as a consequence.

Phase One (Accumulation Period): Under economic pressure, V expands nonlinearly. The supervisor's C_eff degrades, and the error detection rate declines. The penalty does not activate against undetected errors, and feedback to restrain the acceleration of V is absent. Stakeholders perceive that "operations are being conducted safely."

Phase Two (Critical Period): Accumulated errors exceed a critical point as damages in physical space. The materialization of a single incident triggers cascading visibility of errors of the same type. The penalty function fires at a threshold, and V is abruptly suppressed by external forces: litigation, regulatory action, and loss of market credibility.

Phase Three (Reactive Response): Post-incident analysis is conducted, and institutional response commences. However, institutional adaptation has an inherent time delay. By the time processes such as legal reform, updating of organizational protocols, and the formation of standards are completed, technological and market expansion has already entered the next accumulation period in another domain.

 doi:10.20944/preprints202603.1507.v1

Perrow (1984) argued that in highly complex and tightly coupled systems, accidents are not design defects but structural attributes of the system. What this paper demonstrates is that the coupled entity of generative AI and Human-in-the-Loop fulfills precisely the conditions of such a "normal accident" in information space. Complexity derives from the black-box nature of AI output, and tight coupling arises from the permeation of human cognition by AI templates. The very existence of the safety device called Human-in-the-Loop in fact increases the system's complexity and tight coupling, raising the probability of breakdown as a normal accident.

Under these dynamics, what rational choice can decision-makers in high-loss domains make? The next chapter derives the consequence in response to this question.

4. Contraction of Scaling in High-Loss Domains

The preceding chapters have clarified the following mechanism. (1) AI output velocity V expands beyond C_eff, while human cognitive processing capacity C_eff degrades, and institutional adaptation speed Ṡ is bounded. (2) Under this asymmetry, error detection becomes structurally deficient, and gradual adaptation through continuous feedback does not function. (3) Risks accumulate invisibly and materialize discontinuously as threshold shocks. This chapter derives the rational choices available to decision-makers in high-loss domains under this mechanism.

4.1. Divergence of Expected Loss

Generative AI is a probabilistic system. In conventional static risk models, expected loss has been treated as the product "error pass-through rate p × number of trials N × loss magnitude L," with each variable implicitly assumed to be independent of the others. Even if the scale (N) expands, it has been assumed that p is maintained through supervision and that L can be controlled through case-by-case response. In an environment where output velocity V expands, this independence breaks down. An increase in the number of trials raises the effective error pass-through rate p structurally by degrading the supervisor's effective cognitive capacity (C_eff) through cumulative exposure. Concurrently, errors that have accumulated invisibly cause loss magnitude L to inflate discontinuously by materializing as cascading damages the moment the critical point is exceeded. In this context, an increase in V is not merely an increase in N. The expansion of scale itself embeds a circuit that worsens p and amplifies L in a discontinuous manner. Expected loss in high-loss domains is not a product of independent variables but expands through the simultaneous deterioration of interdependent variables.

4.2. Error Reduction Through Capability Improvement

Here we examine the most intuitive objection to this divergence. Namely, the claim that "if AI model capabilities improve, error probability p will approach zero indefinitely, and therefore expected loss should decrease even as V increases." This objection rests on two premises. First, the premise that model capability improvement yields a monotonic reduction in p. Second, the premise that the rate of p reduction exceeds the rate of V increase. Regarding the first premise: it is empirically observed that improvements in AI model capability (improvements in benchmark scores, increases in parameter counts, advances in inference methods, etc.) reduce the error rate on specific tasks. However, the error probability p that concerns us is not the output accuracy of the AI model in isolation but the ultimate error oversight rate in the collaborative work between human and AI. The more the model's output accuracy improves, the more supervisors trust the output and reduce the depth of their scrutiny. The higher the model's performance, the more the checking function by humans becomes a hollow formality, and the rate of C_eff degradation accelerates. Therefore, improvement in AI model accuracy does not necessarily reduce p of the coupled system. Regarding the second premise, even if p of the coupled system decreases slightly, the increase in V follows a nonlinear function, whereas the decrease in p is asymptotic. Because generative AI is a probabilistic system, p = 0 is never achievable. Xu, Jain & Kankanhalli (2024) formally proved, within the

 doi:10.20944/preprints202603.1507.v1

framework of learning theory, that it is impossible for LLMs to learn the totality of computable functions, and therefore that hallucinations cannot be eliminated in principle through scaling of architecture or data. That is, the unattainability of $p = 0$ is not an empirical conjecture but a consequence of computational theory. Consequently, however small p may be, under sufficient increase in V, $p \cdot N \cdot L$ diverges. The linear intuitive understanding that "safety improves as AI model performance improves" does not hold in the face of the non-stationarity of p(V) and the nonlinear expansion of V.

4.3. Flow-Rate Limitation as Rational Equilibrium

Within the framework of this paper, only one among the three constraints is an endogenously modifiable variable. R is embedded in the constitutive principles of the legal system, and C_max is subject to biological constraints. The only operable variable is V.

V_eff = min(V, C_max)

That is, externally limiting AI output velocity to a level that does not exceed the effective human processing capacity becomes the sole policy variable for keeping expected loss bounded. This conclusion is neither a normative nor an ethical claim. Given the premises of the inevitability of probabilistic error, the dynamic degradation of supervisory capacity, and the finiteness of institutional adaptation speed, the problem of expected-loss minimization converges to the equilibrium of flow-rate limitation. Therefore, control of output scale constitutes the rational choice for avoiding structural divergence.

4.4. Capability Improvement Contracts Usage

The consequence derived from the arguments of this paper is clear. The more model capability improves and output velocity V and scope of application expand, the more the use of AI as an autonomous decision-making agent in high-loss domains contracts. Capability improvement accelerates V. The increase in V does not merely increase the number of trials but degrades C_eff through the cognitive adaptation of supervisors, and further promotes threshold-type risk accumulation. As a result, capability improvement does not necessarily reduce expected loss; rather, it can structurally push it upward. Decision-making agents in high-loss domains minimize expected loss under this divergent structure. The rational choice then converges to the limitation of V. In the highest-risk domains where loss magnitude is maximal (medical judgments directly affecting human life, strategic decisions in national security, areas requiring non-probabilistic consistency such as aerospace engineering and nuclear reactor design), the autonomous involvement of probabilistic generative systems converges toward institutional exclusion. This consequence is not a future prediction but an intrinsic implication of our model. And actual institutional design is already moving in the same direction. For example, the EU AI Act introduced use restrictions according to risk classification, effectively establishing flow-rate constraints on high-risk uses. In the medical field as well, AI-assisted diagnosis is subject to verification requirements equivalent to those of existing clinical trials, and the free expansion of output is institutionally restrained. What is important is that this contraction is an equilibrium logically derived from three constraints: the inevitability of probabilistic error, the plasticity of human cognition, and the finiteness of institutional adaptation speed.

5. Limitations and Scope

The model in this paper is not intended as a general theory applicable to all forms of AI use. Here we specify the space of application.

5.1. Domain Limitation

The premise of this paper is that generative AI is a probabilistic system and that error probability p cannot be reduced to zero in principle. This premise holds in domains where the correctness of

 doi:10.20944/preprints202603.1507.v1

output cannot be determined deterministically in advance. In domains such as mathematical theorem proving, program specification verification, and closed physical simulation, it is possible to construct a layer that mechanically judges the correctness of output. If a design is viable that separates a probabilistic generation layer from a deterministic verification layer, as in so-called neuro-symbolic architectures, then it is theoretically possible to bring the effective p close to zero. In high-uncertainty domains such as advanced medical judgment, diplomacy and national security, economic policy, judicial decisions, and corporate management, the correct answer cannot be formalized in advance. The physical world and social space are not closed systems, and the input model itself is incomplete. Unless a verification function exists, the errors of probabilistic generation cannot be systematically eliminated. Accordingly, the scope of this paper is limited to the following domain:

The structural limitations of Human-in-the-Loop generative AI use in decision-making domains where formal verification is inherently impossible and loss magnitude is high.

5.2. Static Character of the Model

This paper is a static argument that discusses the structural consequences derived from the simultaneous satisfaction of the fundamental constraints ($R = 1$, C_{max}, $V \to \infty$). That is, it takes the form of a conditional proposition: "if these conditions hold, then the following consequences are derived." This paper does not quantitatively simulate the specific temporal evolution paths of $V(t)$ or $C_{eff}(t)$. The identification of empirical values for parameters such as degradation coefficients, institutional adaptation speed, and acceleration constants falls outside the scope of this paper. Empirical estimation of these parameters remains a future research agenda as a dynamic extension of this model. All figures presented in this paper are theoretical illustrations. The specific curves, thresholds, and temporal evolutions depicted are conceptual and are not derived from empirical simulations.

5.3. Response to Relative Comparison with Humans

The most classical and powerful objection to this paper's model is one based on relative comparison with human error rates. Namely, the claim that "real humans (doctors, judges, pilots) also commit fatal errors at a certain error probability p_{human}. If the AI error probability p_{AI} is statistically shown to clearly fall below p_{human}, would it not be the case that delegating to AI yields lower expected loss?" This objection rests on two implicit premises. First, the premise that p_{AI} and p_{human} are comparable under identical conditions. Second, the premise that the introduction of AI does not affect p_{human}. The measurement of p_{AI} is typically conducted under controlled evaluation environments (standardized datasets). However, decision-making in the high-uncertainty domains that this paper addresses takes place outside of such controlled conditions. Under conditions involving unknown variables, incomplete information, and context-dependent judgment, p_{AI} may differ from values measured in benchmark environments. In particular, in the aforementioned domains where formal verification is impossible, it is practically difficult to know the true value of p_{AI} in advance. As this paper pointed out in Chapter 2, individual empirical studies have confirmed that the introduction and operation of AI does not hold p_{human} constant. The degradation of supervisory capacity through repeated exposure to AI output may create new cognitive vulnerabilities that would not have arisen had humans been performing the tasks on their own. Furthermore, even if $p_{AI} < p_{human}$ is statistically established, the problem of legal and social liability attribution is not resolved.

5.4. Other Limitations

The model in this paper does not explicitly address several important elements. First, heterogeneity across domains. While this paper discussed "high-loss domains" as a single category, the nature of losses, the institutional frameworks of liability attribution, safety culture, and regulatory environments vary greatly even within fields such as medicine, law, finance, and national security.

 doi:10.20944/preprints202603.1507.v1

Applying this model concretely requires empirical research and institutional analysis specific to each domain. Additionally, this paper presupposes a rational decision-maker, but in real organizations, factors such as the pursuit of short-term profits, competitive pressure, information asymmetry, and intra-organizational politics can produce irrational choices. Even if V limitation is the theoretical equilibrium, whether it is actually adopted depends on organizational governance and market conditions. For example, there will be enterprises that accept a 0.1% probability of a serious AI incident entailing enormous compensation, reasoning that they can offset the loss with the efficiency gains obtained in the 99.9% of cases. Next, the scope of this paper is limited to high-loss domains. It does not deny the utility of generative AI in domains where damage is limited even if errors occur, such as document drafting, ideation support, and preliminary data organization. Where the loss structure differs, the rational equilibrium also differs. The equilibrium analysis in this paper's model presupposes an environment in which the penalty function g(E) operates endogenously through institutional feedback. In domains where the institutional connection between the attribution of damages and the decision-making agent is strong (civil liability, corporate governance, professional responsibility in medicine and law, etc.), the model's predictive power is suggested to be high. On the other hand, in domains such as national security and military operations, where the threshold of g(E) is set exogenously by political judgment, the same structure exists but the conditions for penalty activation depend strongly on the institutional environment and political dynamics, and one should be more cautious in directly applying the equilibrium analysis of this paper.

5.5. Implications for Practical Governance: From Supervision Enhancement to Flow Design

The theoretical consequence of this paper diverges from the design of actual AI governance institutions in certain respects. Current institutional design still implicitly treats Human-in-the-Loop as its foundation, and the prescription for incidents consistently takes a supervision-enhancing approach. These include, multi-layered approval processes (mandatory double-checking), expanded audit logs, augmented checklists, and the imposition of additional supervisory duties on on-site monitors. However, as the model in this paper demonstrates, in an environment where output velocity V expands, measures that increase the burden on supervisory layers are suggested to potentially be counterproductive as a prescription. The practical response derived from this paper's structural analysis is a "flow-design" approach. As the theoretical consequence $V_{eff} = \min(V, C_{max})$ indicates, the true institutional response lies not in forcing human supervisory effort to keep pace with system scale, but in forcibly subordinating the system's output scale to the physical and cognitive limits of humans. Indeed, a similar recognition is beginning to appear in the direction of current institutional design. Bignami et al. (2025) argued that for high-risk clinical uses within the framework of the EU AI Act, proportional allocation of human oversight (adjustment of supervisory levels according to risk tiers) is indispensable, and pointed out that uniform human review across all outputs is practically infeasible. Building on the theoretical foundation we propose, we provide some examples of specific governance measures that should be implemented in high-loss domains:

Hard Caps on Case Loads: Setting a systemic upper limit (hard limit) on the absolute number of AI-assisted tasks that a single supervisor (a doctor, judge, legal officer, etc.) may process and approve per day.

Prohibition of Batch-Approval Architectures: At the UI/UX design level, eliminating at the systems level the functionality for batch approval of multiple AI-generated items, and enforcing serial cognitive load.

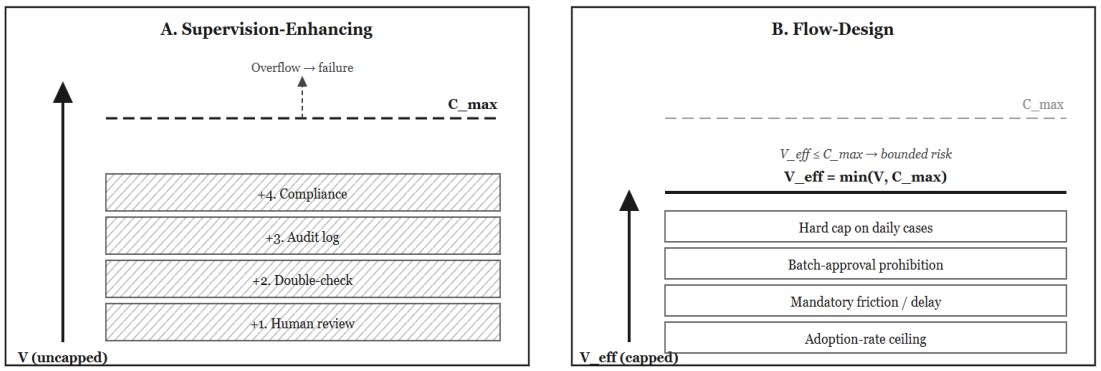
Introduction of Intentional Friction: Incorporating a mandatory temporal delay between the confirmation of AI output or AI-generated products and the point at which the approval button becomes active, intentionally disrupting reflexive approval caused by high processing fluency.

Limitation of Output Density and Adoption Rate: Preemptively limiting the proportion at which AI can autonomously intervene in specific business processes (adoption-rate cap) and the output density per unit of time.

 doi:10.20944/preprints202603.1507.v1

The supervision-enhancing approach, as represented by the EU AI Act among others, and the flow-design approach presented in this paper are aligned in the objective of ensuring safety. However, whereas the former seeks to restrain risk through the addition of supervisory processes and the strengthening of procedural obligations, this study has argued for the necessity of institutionally limiting output velocity itself. The difference between the two lies in the direction of institutional design: whether to respond by increasing supervisory layers, or by physically limiting the flow rate.

Figure 3. Two Paradigms: Supervision-Enhancing vs Flow-Design



(A) adds supervisory layers but cannot exceed C_max; (B) constrains V itself to maintain effective oversight.

Figure 5. A theoretical comparison of two institutional responses to AI scaling. Approach (A) attempts to mitigate risk by adding layers of human supervision; however, this approach ultimately fails as the uncapped output velocity (V) irreparably exceeds the biological cognitive ceiling (C_max). Conversely, the proposed Flow-Design approach (B) preemptively caps the effective output volume (V_eff) to remain strictly within human processing limits, thereby bounding expected loss.

6. Conclusion

In this paper, we have demonstrated that the implicit premise of current AI governance, that human oversight through Human-in-the-Loop guarantees safe scaling, is structurally untenable. The reason lies in the simultaneous satisfaction of three constraints. Liability ultimately rests with humans or legal persons, human cognitive processing capacity has an upper bound, and economic pressures continue to accelerate output velocity. So long as these three conditions coexist, output will eventually exceed human processing capacity, and oversight will degenerate into a formal procedure. Moreover, this hollowing-out is not automatically repaired. Supervisory capacity declines through repeated exposure, and institutional adaptation involves time and friction. As a result, error detection does not function continuously; adjustment is triggered only after a critical shock. The assumption of gradual self-correction does not hold. Under this structure, the rational equilibrium reached by decision-making agents in high-loss domains is clear: limiting output scale to a level that does not exceed the human processing limit. Capability improvement does not automatically enhance safety; rather, the more capability and speed expand, the more the use of AI as an autonomous decision-making agent institutionally contracts. The criterion that this paper presents for practical governance is simple. A design that continues to expand output scale on the premise of oversight embeds not continuous adaptation but threshold-type breakdown. Pre-emptive flow-rate limitation in high-loss domains is not excessive caution but a logical consequence derived from structural constraints. Current AI governance relies on a design philosophy centered on the strengthening of oversight (Human in the Loop). The analysis of this paper shows that oversight strengthening does not form a sustainable equilibrium, and proposes flow design as an alternative institutional axis. The two may coincide in surface-level institutional design and AI governance design in practice, but their theoretical foundations differ.

 doi:10.20944/preprints202603.1507.v1

Declarations

Funding: This study was self-funded by UTIE Instruments Inc. No external grants or third-party funding were received.

Ethics Approval and Consent: Not applicable. This study involves no human subjects, clinical trials, or sensitive personal data.

Data Availability: No proprietary datasets were generated or analyzed.

AI Use: AI-assisted translation, formatting figures. All ideas, discussion, and conclusions are the authors' own.

Competing Interests: The author is the CEO of UTIE Instruments Inc., which provides AI governance advisory services. However, this research was conducted independently as part of the UTIE Research Institute's theoretical framework, and no specific commercial products are promoted.

References

1. Alter, A. L. & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. Personality and Social Psychology Review, 13(3), 219–235. https://doi.org/10.1177/1088868309341564
2. Asgari, E., Montaña-Brown, N., Dubois, M., et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine, 8, 274. https://doi.org/10.1038/s41746-025-01670-7
3. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8
4. Beck, U. (1986). Risikogesellschaft: Auf dem Weg in eine andere Moderne. Suhrkamp. [English edition: Risk Society: Towards a New Modernity, trans. M. Ritter, Sage, 1992.]
5. Bignami, E. G., Russo, M., Semeraro, F., et al. (2025). Balancing innovation and control: The European Union AI Act in an era of global uncertainty. JMIR AI, 4, e75527. https://doi.org/10.2196/75527
6. Carnat, I. (2024). Human, all too human: Accounting for automation bias in generative large language models. International Data Privacy Law, 14(4), 299–314. https://doi.org/10.1093/idpl/ipae018
7. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477
8. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87–114. https://doi.org/10.1017/S0140525X01003922
9. Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260
10. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787
11. Horowitz, M. C. & Kahn, L. (2024). Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. International Studies Quarterly, 68(2), sqae020. https://doi.org/10.1093/isq/sqae020
12. Kahneman, D. (1973). Attention and Effort. Prentice-Hall.
13. Kücking, F., Hübner, U., Przysucha, M., et al. (2024). Automation bias in AI-decision support: Results from an empirical study. Studies in Health Technology and Informatics, 317, 298–304. https://doi.org/10.3233/SHTI240871
14. McCormick, S. (2025). Interpretive debt: How high coherence AI reshapes human judgement, authority, and accountability. SSRN Working Paper. https://doi.org/10.2139/ssrn.5990154
15. Naito, H. (2025). "AI Selection Pressure: Template Saturation and the Reshaping of Human Discernment." Zenodo. https://doi.org/10.5281/zenodo.17644956
16. Omar, M., Brin, D., Glicksberg, B. & Klang, E. (2025). Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine, 5, 171. https://doi.org/10.1038/s43856-025-01021-3

 doi:10.20944/preprints202603.1507.v1

17. Padmakumar, V. & He, H. (2023). Does writing with language models reduce content diversity? arXiv:2309.05196. https://doi.org/10.48550/arXiv.2309.05196
18. Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055
19. Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886
20. Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Basic Books.
21. Warm, J. S., Parasuraman, R. & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. Human Factors, 50(3), 433–441. https://doi.org/10.1518/001872008X312152
22. Xu, Z., Jain, S. & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv:2401.11817. https://doi.org/10.48550/arXiv.2401.11817
23. Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P. & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. arXiv:2409.01754. https://doi.org/10.48550/arXiv.2409.01754
24. [AI & Society review] (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society. https://doi.org/10.1007/s00146-025-02422-7