Supplementary Information

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

Hiroki Naito, UTIE Research Institute (UTIE Instruments Inc.)

S1. The Apparent Effectiveness of HITL Enhancement

S1.1 The Gap Between the Official Narrative and the Actual Mechanism

Currently, the approach taken by many regulatory bodies such as the EU AI Act, as well as legal scholars and policymakers, which holds that strengthening oversight regimes is sufficient, implicitly requires the unrealistic premise that human brain processing speed also grows annually in pace with AI. Nevertheless, with regard to the real world as of 2026, we present the following hypothesis for why a large-scale collapse of human-in-the-loop has not yet surfaced.

The official explanation holds that the strengthening of human oversight improved the quality of supervision, and that accidents consequently decreased. Our explanation, however, is that the strengthening of oversight regimes increased the complexity of procedures, raised the costs of submission, application, and deployment, and thereby reduced V itself. If a five-stage review process exists, it is entirely plausible that submitters would reduce output volume to avoid procedural costs, even if AI can instantly generate plausible documents, and the physical submission constraints from one review body to another certainly reduce speed to some degree. Even if the checks themselves are not functioning effectively, requiring checks creates physical delays, which can exert a suppressive effect. That is, the strengthening of human checking regimes may, in some cases, inadvertently produce effects equivalent to flow design, not through qualitative improvement of supervision, but through the increase of procedural friction.

S1.2 The Time Bomb of Procedural Cost Bypass by Agentic AI

However, the unintended reduction in V carries a time bomb. Once agentic AI evolves to the point where it can automatically generate and submit review documents, procedural costs approach zero. At that moment, if HITL had merely appeared to function by coincidence, its effect vanishes instantly. The flow approach proposed in this paper does not have this vulnerability. For example, if, in addition to identity verification, a physical submission limit of one submission per institution per six months is imposed, it cannot be bypassed regardless of how AI agents evolve. This is because the target of the restriction is not output speed but rather submission count, a discrete, institutional variable.

S1.3. Operational Estimation of C_max

A natural objection is that C_max cannot be measured in advance, rendering the prescription impractical. This section proposes an operational estimation procedure that circumvents the need to measure C_max directly.

The analogy is road speed limits. Speed limits are not derived from measuring each driver's reaction time. They are derived from the physical properties of the road (curvature, visibility, pedestrian density) and set at a level where even the least capable driver remains safe. The estimation targets not the driver's ability but the relationship between speed and accident rate on a given road.

The same logic applies to supervisory capacity in high-loss domains. Rather than measuring the cognitive ceiling of individual supervisors, the procedure estimates the relationship between processing volume and error rate for a given task type, then sets a hard cap below the point where that relationship deteriorates. From existing operational records, outliers (abnormally fast or slow processors) are excluded and supervisors near the median processing speed are selected as a reference population whose performance defines the baseline. Using this reference population's historical records, the relationship between daily case volume and error rate (missed defects, incorrect approvals, post-hoc reversals) is plotted. The hard cap is set not at the point where error rate begins to rise, but at a conservative margin below it. The asymmetry of consequences in high-loss domains justifies this conservatism.

A critical caveat follows from the main text's analysis (Section 2.2). Effective cognitive capacity (C_{eff}) is not static; it degrades over time through cumulative exposure to AI output templates via predictive-error minimization. The error rate measured at the time of calibration will therefore underestimate the error rate after sustained AI-assisted operation. To mitigate this risk, periodic independent audits and red-teaming (deliberate insertion of known errors to measure actual detection rates) should be used to calibrate the baseline against E_{actual}. This procedure requires no new infrastructure. It operates on existing quality-control records and can be implemented within standard operational workflows. The measurement target is not "human cognitive capacity" as an abstract quantity but "the empirical relationship between processing volume and error rate in a specific task domain," which is operationally tractable. Determining precise safety margins and validating this procedure across domains remains a future empirical agenda.

S2. Evaluation of "Thick Oversight" Approaches

In response to the arguments of this paper, there exist approaches whose direction is: "HITL work, let us introduce thicker oversight." These can be classified into three broad lineages. First, the Meaningful Human Control (MHC) enhancement lineage, which seeks to expand meaningful human control over AI decision-making processes. Second, the mandatory Algorithmic Impact Assessment (AIA) and diverse committee review lineage. Third, the due-process (right to contest) and contestability supremacism lineage. All of these are variants of the "supervision-enhancing approach" criticized in Section 5.5 of this paper. They share the common feature of attempting to respond to the diagnosis that HITL is structurally failing by adding oversight layers (committees, third-party audits, appeal procedures). It would not be appropriate to uniformly reject all of these approaches. From the perspective of our framework, a more precise differentiation is possible.

S2.1 Cases Where They Inadvertently Function as Flow Restrictions

These thick oversight approaches can, through the same mechanism as the "unintended reduction

in V" discussed in S1, actually function as flow restrictions in certain situations. If MHC is strengthened, deep human involvement is required for each decision, physically reducing the number of cases that can be processed. If AIA is mandated, the pre-deployment evaluation process may become a bottleneck, constraining the speed of deployment. If contestability is institutionalized, the costs of filing appeals and the waiting time for appeal processing create incentives that indirectly suppress submission and application speed. In all of these cases, supervision quality has not been improved; rather, V has been pushed down through procedural friction. And since procedural friction can be bypassed by the evolution of AI agents, relying on these approaches does not constitute sustainable institutional design.

S2.2 Differences from Intentional Flow Design

The decisive difference between unintended flow restrictions and the intentional flow design proposed in this paper lies in what the restriction targets. The former indirectly pushes down V through procedural costs associated with output speed, but those procedural costs themselves are automatable variables. The latter targets discrete, institutional variables that cannot be bypassed by AI evolution, such as hard caps on submission counts or daily approval limits. This difference determines the durability of institutions in the age of agentic AI. We therefore do not claim that MHC, AIA, or contestability are meaningless. To the extent that they function as flow restrictions, they do in fact contribute to safety. However, one cannot argue that "if it happens to work as a result, even without intent, that is sufficient."

Oversight layers such as committee reviews, third-party audits, and appeal reviews added in a V > C_max environment are operated in a state where the number of cases to be processed exceeds the cognitive processing capacity of human supervisors at each layer. Under these conditions, each oversight layer progressively becomes a hollow formality through the same mechanism discussed in Section 2.2 of this paper: the rewriting of internal standards through repeated exposure to AI templates, the disappearance of a sense of discomfort due to increased processing fluency, and the exhaustion of attentional resources. Hollowed-out oversight layers increase two types of errors. False negatives increase, where inappropriate AI-generated outputs (spurious research, defective legal documents, erroneous clinical judgments, etc.) pass through all oversight layers. So long as the format is well-organized and conforms to plausible statistical regularities, supervisors feel no sense of discomfort and approve without substantive verification of content. Conversely, as discussed in Section 3.2 of this paper regarding the asymmetry of invisible errors, false positives also increase. Even legitimate research or proper applications may be erroneously rejected under the pressure to prioritize rapid processing. That is, the addition of oversight layers contributes to safety improvement as a secondary effect through V reduction via procedural friction, but with regard to oversight quality itself, it poses a risk of increasing both false negatives and false positives.As an illustrative example of this dynamic, consider the submission history of this paper itself. When submitted to a major preprint server, this paper was rejected at the moderation stage (not of interest) and also dismissed upon appeal review. However, the senior administrator responsible for the relevant field at that server expressed the view that the topic of this paper was "obviously of interest." The divergence between the two-stage rejection at the operational level and the senior administrator's judgment is consistent with the AI field on that server having reached

V > C_max due to the surge of AI-generated submissions, and with the increased incidence of false positives under overload conditions.

Finally, in terms of the economic implementation costs of institutions, the two approaches are also asymmetric. The thick oversight approach continuously demands personnel and organizational costs, such as establishing new committees, securing auditors, and operating appeal processing departments, all of which presuppose the additional deployment of humans who are themselves subject to C_max constraints. Flow design can be implemented simply by setting numerical upper limits on existing authentication infrastructure and submission systems, and the maintenance costs of the institution are orders of magnitude lower.

S2.3 Connection to Classical Economics

A concern exists that, even if maintenance costs are low, the introduction of flow design would generally invite competitive disadvantage. However, in inter-platform competition, the opposite dynamic is predicted. On platforms without flow restrictions, the flood of AI-generated content causes the collapse of quality signals maintained by human checking, giving rise to the lemon market problem demonstrated by Akerlof (1970). On platforms that have introduced flow design, the submission restriction itself functions as a quality signal, attracting high-quality submissions and users. Restrictions do not undermine credibility; restrictions constitute credibility. However, the conditions under which this dynamic operates effectively depend on the structure of the penalty function g(E) discussed in Chapter 3 of this paper. In markets such as inter-platform competition, where customers can evaluate quality to some degree and where declining credibility is internalized through user attrition, g(E) operates endogenously, and the superiority of flow design as a quality signal is realized. Already, major video platforms have introduced account suspension measures in response to mass submissions of AI-generated content, which is consistent with reality as an instance of flow design being voluntarily adopted under market pressure.

S3. Comparison of Arguments with Recent Related Research

Green (2022) was among the earliest to identify that human oversight requirements can function as false assurance rather than genuine safety mechanisms. While sharing the starting point and recognition of this paper's HITL critique, their proposed alternatives remain limited to "improving the algorithms themselves" and "strengthening pre-deployment regulation," and do not reach flow design.

Bastani and Cachon (2025) demonstrated the economic impossibility of sustaining oversight incentives. They arrived at an ostensibly similar conclusion during the same period as the writing of this paper. Namely, the claim that the more AI system reliability improves, the more difficult and costly it becomes to economically motivate human oversight. Both papers share the paradoxical claim that "AI capability improvement makes human oversight more difficult," but they differ in their theoretical foundations, the locus of the problem, and the direction of solutions.

The model of Bastani & Cachon is grounded in the principal-agent problem (contract theory). The source of the paradox is the misalignment of economic incentives. As AI becomes more accurate,

errors become rare, making it difficult to design rewards for supervisory effort. The collapse mechanism is that supervisors rationally shirk (moral hazard), and no technical or cognitive deficiency is assumed. The direction of the solution is the improvement of contract design: the reconstruction of incentive structures commensurate with oversight costs. Our argument, on the other hand, is grounded in the integration of cognitive science, institutional theory, and risk engineering, as already described in the main text. In other words, the two models are complementary. Bastani & Cachon demonstrate that "even if the human oversight regime is fully in place, supervision cannot be economically motivated," while this paper demonstrates that "even if economically motivated, the human oversight regime itself becomes structurally untenable." If both hold as facts, then the improvement of contract design alone is insufficient, and the conclusion of this paper, that institutional restriction of output volume itself is the only rational equilibrium, is more strongly supported.

Yang & Zhu (2026) quantitatively demonstrated a quality-dependent bias in which AI peer review overvalues low-quality papers and undervalues high-quality papers. They also reported the fact that editors are informally discounting AI reviews. These findings provide valuable empirical evidence that quality signals in peer review are already degrading under AI integration, corroborating the structural argument of this paper. Their empirical methodology offers a foundation for measuring signal degradation across domains, though the theoretical connection to flow design remains to be developed.

Htin (2026) analyzed human oversight through the Learned Hand negligence formula ($B < P \times L$), concluding that mere human presence is insufficient and identifying the moral crumple zone in which human operators absorb liability for systemic design failures. The diagnosis aligns closely with the present paper. However, the proposed remedy, a three-pillar framework imposing duties of genuine human-AI collaboration, technical robustness, and post-market monitoring, remains within the supervision-enhancing paradigm.