Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Control Assimilation in Agentic AI: Formal Human Authority Persists as Operational Access to Control Becomes Technically Mediated

Christian Jörg Rustenbach

christian.rustenbach@med.uni-tuebingen.de

Universitätsklinikum Tübingen https://orcid.org/0000-0003-2178-8397

Christoph Salewski

Universitätsklinikum Tübingen https://orcid.org/0000-0001-7155-6039

Annika Kral

Otto-von-Guericke University Magdeburg

Stefan Reichert

Universitätsklinikum Tübingen

Research Article

Keywords: agentic AI, human control, human-in-the-loop, autonomous systems, governance, control assimilation

Posted Date: September 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11160733/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background: “Human-in-the-loop” is often treated as evidence of control, yet it does not establish which information reaches the human, when intervention becomes possible, what actions remain executable, or whether the route to intervention is technically governed.

Methods: We performed a prespecified, cross-sectional, provenance-preserving architecture-level evidence synthesis of operationally described real-world or deployment-proximal agentic and highly autonomous AI-enabled systems. Searches across Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and OpenAlex yielded 6,754 unique canonical Work IDs. A frozen priority-routing scheme assigned 1,156 records to the first-wave P1 cohort; 990 underwent substantive full-text assessment. Evidence from 25 retained source/manifestation Work IDs was consolidated into 17 heterogeneous architecture cases. We coded control-function allocation, responsible-human role paths, technical mediation, intervention conditions, formal authority, control mechanisms, and documented effective-control consequences. After taxonomy freeze, an independent blinded human reviewer reclassified standardized neutral summaries of all 17 architectures.

Results: Eleven architectures were RC-A, four RC-B, and two RC-C. Local veto/override was more strongly documented than global/system stop in 16/17 cases. System- or policy-mediated recruitment of human attention was clear in 14/17, mixed in one, and indeterminate in two. Among 14 outcome-evaluable architectures, direct human control exercise was documented in 5/14 under the base rule and 3/14 under a strict high-confidence rule. Locked-rule reclassification reproduced all 17 labels (Cohen’s $\kappa = 1.00$).

Conclusions: Across this frozen cohort, formal human authority often persisted while operational access to control was technically mediated. We term this structural condition Control Assimilation. This distinction concerns governance and institutional oversight.

1. Introduction

Human oversight has become a central safety and governance requirement for artificial intelligence. The European Union Artificial Intelligence Act, for example, requires high-risk AI systems to permit effective oversight by natural persons and explicitly refers to the ability to understand system limits, monitor operation, disregard or reverse outputs, intervene, and stop the system (European Parliament & Council of the European Union, 2024). Yet the phrase 'human-in-the-loop' remains structurally underspecified. A human may be present somewhere in a workflow without controlling what information is surfaced, which cases are escalated, when an intervention window opens, or whether the available action can still alter the system trajectory. Recent work on agentic AI has therefore moved from simple checkpoint metaphors toward layered agency, adaptive intervention, and explicit oversight architectures (Zhu et al., 2026; Kumar & Singh, 2026).

This problem has deep roots in human-factors research. Bainbridge's 'ironies of automation' showed that automation can remove the human from routine control while leaving the person responsible for

abnormal situations that are intrinsically difficult to manage (1983). Parasuraman, Sheridan, and Wickens subsequently formalized automation as a function-allocation problem spanning information acquisition, analysis, decision/action selection, and action implementation (2000). Endsley and Kiris demonstrated the out-of-the-loop performance problem, in which operators supervising automation can lose situation awareness and become slower or less effective when manual takeover is suddenly required (1995). Later synthesis linked automation-related complacency and automation bias to the allocation of human attention itself (Parasuraman & Manzey, 2010). Together, these literatures established that automation changes human work rather than merely reducing it.

Contemporary accounts of Meaningful Human Control make the normative implication explicit: responsibility should be commensurate with a person's actual ability and authority to control the system, not merely with formal role assignment (Santoni de Sio & van den Hoven, 2018; Cavalcante Siebert et al., 2023). Green's analysis of human-oversight policies similarly warns that nominal human review may create false assurance when humans cannot reliably perform the expected corrective function (2022). Langer, Baum, and Schlicker operationalize a complementary part of the problem through signal-detection theory, emphasizing reliable error detection as a crucial condition for effective human oversight (2025). The classic responsibility-gap problem sharpens the same concern from another direction: increasing system autonomy can separate responsibility for outcomes from the actors who can predict or causally govern them (Matthias, 2004). Human-AI interaction research has consequently focused on calibrated reliance, contestability, and interaction designs that force active evaluation rather than passive acceptance (Amershi et al., 2019; Bucinca et al., 2021; Mehrotra et al., 2024).

Agentic AI introduces a further architectural problem. An agent may not merely generate a recommendation for human consideration; it may observe, classify, prioritize, plan, call tools, execute actions, monitor consequences, and decide when to recruit a human. Under these conditions, the human-control pathway itself can become a product of technical processing. The system or surrounding governance infrastructure may determine which event is considered exceptional, which representation is shown to the operator, which option set is available, and how much residual time remains for intervention. Takahashi's oversight-centered metrology independently formalizes this concern by treating humans and automated verifiers as costly interrupt channels subject to routing error, delay, availability, congestion, and irreversibility constraints (2026). Formal authority can therefore persist while practical access to control becomes technically conditioned.

We use the term Control Assimilation for this architecture-level divergence. Control Assimilation does not mean that humans disappear, that all autonomy is unsafe, or that technical mediation necessarily defeats human agency. It denotes a more specific structural condition: control functions that shape perception, case selection, authorization, timing, or intervention are incorporated into the autonomous technical architecture, such that a human may retain formal or terminal authority while reaching that authority through a technically mediated runtime control path. Crucially, this mediation need not be performed by a second AI model or supervisory agent: deterministic policy engines, thresholds, software gates, or other governance infrastructure can mediate the control route when they technically determine what reaches

the human, when the human is recruited, or what remains executable at that point. In this study, 'assimilation' is used structurally rather than longitudinally: our cross-sectional architecture evidence does not show that any individual system became progressively more assimilated over time. The construct therefore differs from a simple level-of-automation scale, from automation bias as a cognitive effect, from Meaningful Human Control as a normative requirement, from Nguyen et al.'s taxonomy of controllability paradigms (2026), from the supervision-to-teaming account of operational human control described by Tsamados et al. (2025), and from Takahashi's metrology of oversight-channel performance (2026). Its empirical object is the route to human control itself: where the human control boundary is located, which machine or policy transformations occur upstream, and what remains executable when that boundary is reached.

We addressed this question using a prespecified, provenance-preserving synthesis of operationally described real-world or deployment-proximal agentic and highly autonomous systems across domains. The empirical unit was a concrete architecture case supported by implementation- or deployment-context evidence, not a paper count or a catalogue of purely proposed oversight concepts. Rather than count papers as independent observations, we reconstructed unique architecture cases and responsible-human control paths. We asked four primary questions: (1) whether human control was global or predominantly local and selectively gated; (2) whether human attention and intervention opportunities were themselves recruited through system or policy routing; (3) whether formal governance architecture corresponded to observed direct human control exercise; and (4) whether recurring configurations could be identified without collapsing heterogeneous systems into a single autonomy or 'control' score. The aim was not to estimate the population prevalence of Control Assimilation, but to determine whether it is a coherent and empirically reconstructable architecture-level phenomenon.

2. Methods

2.1 Study design and prespecification

This study was designed as a provenance-preserving, architecture-level evidence synthesis of human control in agentic or highly autonomous AI-enabled systems. The novelty-search protocol v0.2 was versioned before the subsequent platform-specific completion runs, and retrieval amendments were prospectively documented when provider limits required partitioning or hybrid treatment. Search retrieval, deduplication, novelty screening, empirical architecture eligibility, full-text acquisition, and analytic coding were maintained as distinct stages. The target construct was Control Assimilation: a structural relocation of operational control functions into autonomous or agentic technical architecture such that human oversight may be indirect, selective, supervisory, escalatory, or technically mediated. The term is descriptive of architecture state in the present cross-sectional evidence base and is not used as evidence of a temporal transition within an individual system. The review therefore did not target the broader human-in-the-loop literature indiscriminately. Eligible evidence had to be informative about operational control architecture, including function allocation, intervention or backstop mechanisms, machine-mediated observation or selection, or observed consequences of human or independent control.

The 17-case analytic denominator and all case-level coding were determined exclusively from the frozen review evidence base. Foundational and contemporary literature cited in the Introduction and Discussion was used for conceptual comparison and interpretation and did not alter inclusion, classification, or outcome coding.

2.2 Information sources and search strategy

The reproducible core search stack comprised Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and OpenAlex. Web of Science was executed in the authenticated institutional environment; IEEE Xplore and ACM Digital Library were executed with platform-specific, versioned query adaptations where provider query or export limits required partitioning; and OpenAlex was executed as a reproducible cross-domain complement. The executed search log preserved raw hit counts, query strings, platform adaptations, export limitations, and retrieval-completion status rather than treating provider constraints as zero results. Google Scholar and Semantic Scholar were used for discovery and citation chaining rather than as sole reproducible databases. arXiv, OpenReview, publisher pages, official technical sources, institutional repositories, Zenodo, and related repositories supported frontier-literature discovery, manifestation resolution, and lawful exact-work acquisition. Crossref was used for DOI/metadata validation. Scopus was prespecified as an optional sensitivity source but was not required when independent access was unavailable.

Nine complementary search families were executed or retained under the versioned protocol. S1 targeted broad agentic systems combined with human control or oversight; S2 added machine-mediated oversight/supervision; S3 added control-architecture and competing-construct terms; S4 formed a Meaningful Human Control bridge; S5 targeted AI-control, scalable-oversight, trusted-monitoring, recursive/nested-oversight and related terms; S6 bridged human-factors, supervisory-control, function-allocation, levels-of-automation, attention-demand, out-of-the-loop, neglect-time and fan-out concepts; S7 targeted deployment/production and real-world agent settings; S8 targeted authorization, permissions, sandboxing, containment, defense-in-depth and hard/operational constraints; and S9 performed exact-phrase and close-variant checks for the candidate project terminology. Platform-specific splits or sensitivity refinements were documented as amendments and did not silently replace the original search-family logic.

Query-level result counts were retained separately because overlap between search tracks was substantial and the OpenAlex backfill was not Boolean-equivalent to the primary searches. Consequently, query hit counts were not summed to derive the analytic denominator. The authoritative search denominator was the deduplicated and canonicalized master-work set described below.

2.3 Record canonicalization, deduplication, and title/abstract prioritization

Records were consolidated into a canonical master dataset using a predefined identity hierarchy: normalized DOI first, PMID second where available in imported metadata, and normalized title third. Title plus first-author/year similarity was used only to flag records for manual identity review and was never

sufficient for automatic merging. For every duplicate decision, source identifiers, the keeper record, duplicate identifiers, and the merge rationale were retained. This process yielded 6,754 unique canonical Work IDs. At the final canonical checkpoint, database-provenance tags were present for 6,436 Work IDs from OpenAlex, 525 from Web of Science, 514 from IEEE Xplore, and 45 from ACM Digital Library; these source counts overlap and are therefore not additive denominators. Query-level hit counts were likewise not pooled because search families and database coverage overlapped substantially.

Title/abstract prioritization identified records most likely to support reconstruction of an operational human-machine control architecture. P1 routing favored plausible AI-enabled, autonomous, or agentic systems with human monitoring, approval, intervention, escalation, override, backstop, or role-allocation pathways; records with lower immediate architecture yield remained in P2 for later full-text work. Pure model benchmarking, generic usability or trust studies without reconstructable operational control allocation, non-AI automation, and purely theoretical ethics were therefore less likely to enter the first-wave lane. Crucially, this was prioritization rather than a final eligibility judgment. The Phase-1 workflow was intentionally sealed at the prespecified first-wave analysis freeze, with P2 remaining OFF unless separately authorized. The present analysis must therefore be interpreted as a prespecified first-wave architecture-level evidence synthesis, not as exhaustive full-text adjudication of all 6,754 canonical records.

2.4 Exact-work full-text acquisition and manifestation control

Full-text acquisition was tracked in an append-only manual acquisition ledger and was kept analytically separate from substantive screening. Identity was resolved at the work and manifestation level: DOI, repository concept DOI, concrete version DOI, journal version, accepted manuscript, preprint, and technical artifact were not silently collapsed. A non-identical sibling record, adjacent repository identifier, later version, or thematically similar paper was not substituted for an inaccessible target unless an authoritative provider or author source explicitly established the relationship. Where a preprint or repository copy was used for the same published work, the manifestation relationship was retained in the provenance record. No missing exact full text was converted into substantive evidence through an OCR proxy or abstract-only inference.

At the prespecified analysis freeze, 990 of the 1,156 P1 records had undergone substantive full-text assessment and 166 had not. The latter group comprised records whose substantive full-text adjudication had not been completed at the freeze. These records were retained as a neutral non-adjudication category rather than treated as substantively excluded or empirically ineligible. Within the 990 substantively assessed records, 753 were labelled near_match_FT and 237 not_near_match_FT for novelty/relevance tracking; these labels did not determine empirical architecture eligibility.

2.5 Empirical architecture eligibility and analytic unit

Substantively assessed full texts were reviewed for whether they supported reconstruction of a concrete operational human-machine control architecture: the allocation of control functions, the human role, the

path by which system state or selected cases reached the human, available intervention or backstop mechanisms, and the implementation or deployment context. Purely conceptual oversight proposals without a reconstructable operational control path were not sufficient for architecture eligibility. Architecture eligibility was adjudicated independently of publication count. Multiple publications, repository manifestations, or versions referring to the same underlying system were therefore consolidated rather than treated as independent observations.

The frozen primary evidence bundle contained 25 unique source/manifestation Work IDs supporting 17 unique architecture cases (AP1A001-AP1A017). The architecture case, not the paper or Work ID, was the analytic denominator. Complete case-to-source mapping, including version and DOI relationships, is reported in Supplementary Table S1. Fifteen architecture cases were classified as digital/software-agentic and two as embodied/physical-agentic; 15 were classified as D1 and two as D2 according to the frozen deployment schema. Fourteen cases were outcome-evaluable; A012, A014, and A017 lacked a sufficiently narrow observed human-control consequence for the outcome layer.

2.6 Control-path coding and construct definitions

Control functions were coded as separable components rather than collapsed into a global autonomy score: observation/filtering, interpretation, prioritization, authorization, execution, intervention, rollback/recovery, and audit. For each responsible-human role, a distinct control path was reconstructed. Across the 17 cases, 31 role paths were identified; 11/17 architectures contained two or more distinct paths.

Minimum machine-mediated control depth (MMCD) was defined as the minimum number of necessary machine-mediated transformations on the path between the relevant system state/event and the responsible human control point. Generic network or software hops without decision-transforming relevance were excluded. Distinct human roles or paths were not averaged. Of 31 role paths, 27 had a numeric MMCD, 25 had a positive numeric minimum, two had a documented zero minimum that did not imply zero overall system depth, and four were non-scalar or numerically indeterminate.

Intervention opportunity (whether a responsible human received a usable intervention point in time), intervention capability (an executable human action), formal human authority (documented allocation of decision or governance authority), and effective control (an observed practical control consequence) were coded separately. Local mechanism states were coded for stop, veto, override/takeover, rollback, and recovery. UNKNOWN denoted insufficient evidence and was never converted to zero or absence; internally inconsistent evidence was retained as DISPUTED rather than forced to resolution.

The exact human-review proportion (OCR) required a same-universe denominator and an evidenced numerator of completed, meaningful human review. Routing, display, logging, notification, escalation, or a requirement for approval did not count as completed review. No architecture met the requirements for an exact OCR estimate, and no proxy review-rate comparison was performed.

Independent Constraint Criterion (ICC), Variant B, required an enforcement mechanism whose operation did not depend on voluntary model compliance and that the acting model/path could not disable, bypass, or modify at enforcement time. SUPPORTED_SCOPE_LIMITED was used when such independence was demonstrated only for a documented load-bearing boundary rather than across the entire system.

2.7 Phase-2 configuration synthesis and counterexample testing

Phase 2 was conducted only after the 17-case population had been frozen. Construct-wise distributions and path-level descriptions were examined before configuration labels were assigned. Recurring configurations were then derived by explicit decision rules and tested against all 17 cases, including strongest-alternative and counterexample review.

RC-A (selective gate-based delegation) required at least one clear selected pre-action, pre-effect, or pre-change human gate coexisting with routine delegated or autonomous execution. RC-B (exception/takeover control) captured architectures in which human control was primarily recruited through failure, exception, monitoring, transfer, or residual override without a clear universal timely pre-action gate. RC-C (end-user final-action control after machine filtering) was reserved for architectures in which the responsible human was an end user who retained final accept/reject/appeal or physical-action authority after upstream machine filtering, ranking, framing, or suppression. For RC-C, end-user/final-action role had explicit precedence over generic gate status. The governance-over-epistemic/evolution boundary (O1) was coded as a cross-cutting overlay rather than a fourth primary configuration.

All 17 assignments underwent a final adversarial classification stress test against the strongest plausible alternative configuration. No membership changed. A003 and A012 remained RC-B with MODERATE membership confidence because trigger/timing evidence was incomplete; all other memberships were HIGH. RC categories were treated as descriptive recurring configurations, not latent natural classes, an ordinal autonomy scale, or a causal exposure.

After the RC decision tree and all 17 primary assignments had been frozen, classification reproducibility was tested by an independent reviewer who had not participated in taxonomy derivation. The reviewer received 17 standardized neutral architecture summaries in randomized order (R01-R17) together with the locked three-step RC decision tree. Original architecture identifiers, original RC labels, outcome classifications, manuscript text, and source-paper identities were withheld. The reviewer was instructed to classify from the supplied cards only and not to use external literature, web search, the study manuscript, or generative-AI assistance. For each case, the packet requested an RC label (RC-A, RC-B, RC-C, or UNCERTAIN), a confidence rating (HIGH, MODERATE, or LOW), and a decisive-criterion field. The returned workbook was cryptographically frozen before comparison with the original labels. RC disagreements, if present, were prespecified to be preserved before any adjudication; no consensus recoding was permitted before agreement statistics were calculated.

2.8 Outcome-evidence layer, stratification, and sensitivity analyses

Outcome analysis was restricted to the 14 outcome-evaluable architectures. Evidence was classified conservatively as: (1) high-confidence direct human exercise observed; (2) moderate-confidence direct human exercise, where the human act was explicit but narrow/test-specific or architecture-level attribution was limited; (3) an operational control effect observed without separable attribution of the human contribution; or (4) indirect or primarily machine-control outcome relevance without a narrow exercised human-control consequence.

Deployment class (D1 versus D2) and system stratum (digital versus embodied) were examined separately before any pooling. Because RC-B was compositionally entangled with D2/embodied cases and the embodied outcome-evaluable stratum contained only one case, cross-stratum causal or comparative pooling was not authorized.

Sensitivity analyses restricted the population to outcome-evaluable cases, Digital+D1 cases, E1_DIRECT evidence, and HIGH-confidence RC memberships; tested a counterfactual gate-first classification rule; compared supported-only with supported-plus-partial mechanism coding; compared full versus scope-limited ICC coding; and repeated direct-human-exercise coding using only HIGH-confidence direct cases. All sensitivity results and the adversarial classification stress test are reported in Supplementary Table S4.

2.9 Statistical and interpretive analysis

Scientific analyses were descriptive. Counts and within-sample proportions characterized architectures and mechanism states. We did not perform hypothesis tests comparing RCs, deployment classes, mechanism states, or outcome-evidence categories because the 17 architecture cases were heterogeneous, cell counts were small, and several candidate comparisons were structurally non-independent from the configuration rules or deployment composition. In this setting, a nominal p-value could quantify taxonomy construction or sample composition rather than a population effect; robustness was therefore evaluated through explicit restriction, recoding, adversarial sensitivity analyses, and independent reclassification instead. One exploratory workflow diagnostic was used for attrition assessment: substantive full-text adjudication status (assessed versus not assessed) was compared across the already fixed P1A/P1B/P1C priority strata using Pearson's chi-square test, with Cramér's V as an effect-size descriptor. This diagnostic addressed whether non-adjudication was visibly concentrated in one workflow stratum; it was not treated as a scientific-effect test and was not interpreted as evidence that missingness was random or unbiased. Independent RC reproducibility was summarized using exact percent agreement, the three-category confusion matrix, and Cohen's kappa; because RC-C was uncommon, kappa was interpreted together with raw agreement rather than in isolation. Confidence agreement was reported descriptively. No regression, causal effect estimate, configuration ranking, composite Control Assimilation score, or imputed OCR analysis was performed.

2.10 AI-assisted research workflow and human accountability

OpenAI ChatGPT, including GPT-5.6 Sol and GPT-6 Astra (OpenAI), was used to support literature discovery and search-strategy refinement, evidence organization and consistency checks, and exploratory

calculations. All source-based claims and relevant outputs were verified against the underlying evidence or authoritative metadata. Study eligibility, evidence adjudication, dataset freezes, final classifications, scientific interpretation, and conclusions were determined by the human authors, who retain full responsibility for the submitted work. The independent blinded RC reclassification was performed by a human reviewer without generative-AI assistance.

3. Results

3.1 Study flow and analytic population

The canonical search and reconciliation workflow yielded 6,754 unique Work IDs. Frozen priority routing assigned 1,156 records to the first-wave P1 lane (P1A n = 172; P1B n = 85; P1C n = 899) and retained 5,598 records in P2 as indeterminate_requires_FT; P2 was not substantively excluded and was not advanced before the analysis freeze (Fig. 1). Within P1, 990 records underwent substantive full-text assessment and 166 did not. Non-adjudication was 12.8% in P1A (22/172), 17.6% in P1B (15/85), and 14.3% in P1C (129/899), with no detectable concentration across the three priority strata in the exploratory workflow diagnostic ($\chi^2(2) = 1.09$, $p = 0.579$; Cramér's $V = 0.031$). This finding does not establish that non-adjudication was random or unbiased. Architecture-eligible evidence was consolidated from 25 retained source/manifestation Work IDs into 17 unique architecture cases; 14/17 were outcome-evaluable. The 17-case population comprised 15 digital/software-agentic and two embodied/physical-agentic systems, with 15 D1 and two D2 deployments (Table 1; Supplementary Tables S1-S3).

3.2 Three recurring locations of the human control boundary

Eleven of 17 architectures (64.7%) were classified as RC-A, selective gate-based delegation: A002, A005-A008, A010-A011, A013-A015, and A017 (Pratinav & Kataria, 2026; Sharma, 2026; Malik et al., 2026; Vinjamuri, 2026; Chalupka, 2026; Vangoor, 2020; Ravi & Patlolla, 2024; Chittala & Unnava, 2026; Sweet, 2026; Ravi Kumar, 2026; Idris, 2026). Four (23.5%) were RC-B, exception/takeover control: A001, A003, A004, and A012 (Yu et al., 2026; Maheswaran et al., 2010; Mukherjee et al., 2026; Ahmed, 2025). Two (11.8%) were RC-C, end-user final-action control after upstream machine filtering: A009 and A016 (Zhang et al., 2026; Teixeira, 2026) (Fig. 2; Table 1). RC-A and RC-C assignments were unchanged by the final adversarial classification stress test; A003 and A012 remained RC-B with MODERATE confidence because the timing or trigger of usable human intervention was incompletely specified (Supplementary Table S4). These proportions are descriptive of the frozen analytic cohort and are not estimates of population frequency or prevalence.

RC-A architectures positioned the human at selected pre-action, pre-effect, promotion, policy-change, or high-risk authorization boundaries while routine subsets remained delegated. RC-B architectures positioned human control primarily at failure, exception, transfer, monitoring, or residual override pathways. RC-C preserved terminal user agency, but the system had already shaped the information or option set through filtering, ranking, framing, or suppression. The gate-first counterfactual moved both RC-

C cases into RC-A, confirming that RC-C depends on the explicitly prespecified role/final-action precedence rule rather than representing an independent latent class.

3.3 Human control was predominantly local and selectively recruited

Documented global/system stop capability was substantially less common than local control mechanisms. Global or system-level stop was supported in one architecture (A005 (Sharma, 2026)), partial in one (A014 (Sweet, 2026)), and unknown/not documented in 15. By contrast, veto was supported in 10/17 and partial in 2/17, while override/takeover was supported in 8/17 and partial in 5/17. Under the prespecified evidence-state coding, at least one local veto or override/takeover mechanism was more strongly documented than global/system stop in 16/17 architectures and tied it in A005; no architecture had stronger documentation for global stop than for its strongest local mechanism. This is explicitly a comparison of documentation strength, not a capability-prevalence estimate: UNKNOWN denotes insufficient evidence and is never interpreted as absence. All counts in this comparison are within-cohort descriptions and do not support population-frequency inference.

Human attention and intervention access were commonly recruited through upstream technical selection. Clear system- or policy-gated attention recruitment was documented in 14/17 architectures, was mixed in A009 (Zhang et al., 2026), and was indeterminate in A003 (Maheswaran et al., 2010) and A012 (Ahmed, 2025). Examples included failure detection and operator allocation in ARMADA (Yu et al., 2026), policy GATE routing in AEGIS (Sharma, 2026), severity/confidence routing in closed-loop cloud remediation (Ravi & Patlolla, 2024), and context/fatigue/validity filtering before user nudges in the energy-intervention system (Teixeira, 2026). Importantly, 'system-gated' did not necessarily mean model-gated: deterministic policy engines, thresholds, class rules, and governance logic also mediated access to the human control point.

3.4 Machine-mediated path structure

At least one positive documented MMCD path was present in all 17 architecture cases. Eleven architectures contained two or more distinct responsible-human role paths. Across 31 role paths, 25 had positive numeric MMCD minima; two had a documented zero path minimum that did not imply zero overall architecture depth, and four were non-scalar or numerically indeterminate. Because empirical architecture eligibility required a reconstructable human-machine control path, the presence of a positive MMCD path in all 17 cases is not interpreted as an independent prevalence estimate of machine mediation; it characterizes the internal structure of the selected architecture cohort. MMCD also did not distinguish RC-A, RC-B, and RC-C sufficiently to justify its use as a configuration classifier or global score (Supplementary Table S3).

3.5 Architecture-level governance was more commonly demonstrated than direct human control exercise

Fourteen architectures were outcome-evaluable. High-confidence direct human control exercise was observed in 3/14 cases: teleoperator takeover/correction in ARMADA (A001 (Yu et al., 2026)), user preview/appeal/reweighting in the recommendation-filtering system (A009 (Zhang et al., 2026)), and final user action after machine-selected energy nudges (A016 (Teixeira, 2026)). Two additional cases were coded as direct under the base rule but only with MODERATE attribution confidence: manual Level-D verification in Machine-Native Intelligence v2 (A008 (Chalupka, 2026)) and owner confirmation/investigation after pipeline drift detection (A015 (Ravi Kumar, 2026)). Thus, the base direct-human-exercise count was 5/14 (35.7%), whereas the strict HIGH-confidence count was 3/14 (21.4%) (Table 2; Supplementary Tables S3-S4).

In seven cases (A002, A004-A007, A010, and A011 (Pratinav & Kataria, 2026; Mukherjee et al., 2026; Sharma, 2026; Malik et al., 2026; Vinjamuri, 2026; Vangoor, 2020; Ravi & Patlolla, 2024)), an operational gate, block, escalation, or other control consequence was observed but the completed human action or its causal contribution could not be isolated. Two further cases (A003 and A013 (Maheswaran et al., 2010; Chittala & Unnava, 2026)) provided primarily indirect or machine-control outcome evidence. These distinctions prevented us from inferring effective human control solely from the presence of governance mechanisms, formal authority, or generic system performance.

ICC Variant-B evidence was documented in 8/17 architectures overall: three SUPPORTED and five SUPPORTED_SCOPE_LIMITED. Within the 14 outcome-evaluable cases, documented ICC evidence occurred in 6/14, all within RC-A. This distribution was not treated as an independent association because gate-based configuration criteria and independently enforced boundaries are structurally related. Restricting ICC to full SUPPORT reduced the RC-A count from 6/9 to 3/9 without changing the direction of the descriptive pattern.

3.6 Deployment and stratum sensitivity

Configuration composition differed markedly by deployment and stratum. Among the 14 outcome-evaluable architectures, all nine RC-A and both RC-C cases were Digital+D1, whereas two of three RC-B cases were D2 and one of three was embodied. Both outcome-evaluable D2 architectures were RC-B and emphasized fallback/takeover rather than a clear pre-action gate. The only outcome-evaluable embodied architecture was ARMADA (A001 (Yu et al., 2026)); the second embodied case, A012 (Ahmed, 2025), was not outcome-evaluable. Accordingly, no independent D1-versus-D2 or Digital-versus-Embodied effect was estimated, and cross-stratum pooling remained unauthorized.

3.7 Robustness of the descriptive synthesis

The RC-A/RC-B/RC-C structure persisted when analysis was restricted to the 14 outcome-evaluable cases (9/3/2). Restricting further to Digital+D1 cases reduced RC-B from three to one, demonstrating material deployment/stratum sensitivity. Restricting to E1_DIRECT evidence yielded RC-A/RC-B/RC-C counts of 7/3/0, so RC-C was retained as a qualitative boundary configuration rather than a robust prevalence class. Excluding MODERATE memberships yielded 9/2/2 in the outcome-evaluable subset. The counterfactual gate-first rule yielded 11/3/0, moving A009 and A016 from RC-C to RC-A. Restricting direct-human-

exercise evidence to HIGH confidence reduced direct cases from 5/14 to 3/14 and removed both RC-A direct cases, while the RC-B and RC-C direct cases remained unchanged. In the separate formal-publication-channel sensitivity set (n = 8; seven outcome-evaluable), local veto/override evidence exceeded the documented global-stop state in 8/8 cases and clear system/policy-mediated human-attention recruitment remained present in 6/8, whereas direct human control exercise was observed in only 1/7 outcome-evaluable cases under both the base and strict attribution rules. RC-C was absent from this restricted set. These analyses preserve the direction of the local-control and mediated-access findings while showing that RC-C frequency and direct-exercise evidence are source-sensitive (Supplementary Tables S4-S5). The frozen RC labels were subsequently subjected to independent blinded reclassification as reported below.

3.8 Independent reproducibility of RC classification

The independent blinded reviewer applied the locked RC decision rule to all 17 standardized neutral architecture summaries. All 17 labels exactly matched the frozen primary classification (17/17; 100% agreement): RC-A 11/11, RC-B 4/4, and RC-C 2/2, yielding Cohen's kappa = 1.00. No RC-level adjudication was required. Classification confidence agreed in 16/17 cases (94.1%). A003 reproduced the original RC-B/MODERATE assignment; A012 was independently classified RC-B with LOW confidence compared with the original MODERATE confidence, preserving the configuration while independently reinforcing the pre-existing uncertainty around trigger/timeliness (Supplementary Table S4).

4. Discussion

4.1 Principal finding: the loop can remain human while access to it is machine-governed

The strongest finding of this synthesis is not that humans were simply removed from control, nor that human control functions were uniformly transferred to a second AI supervisor. In most architectures, formal human authority, approval rights, override capability, or a terminal human action remained visible. The architectural change lay upstream: technical or policy logic often determined which states were detected, which cases were classified as exceptional or high risk, which evidence was compressed and displayed, and when the human was recruited. Clear system- or policy-mediated attention recruitment was documented in 14/17 architectures, while local veto/override mechanisms were more strongly documented than global/system stop in 16/17. The mediating layer could be model-based, but it could also consist of deterministic thresholds, policy rules, software gates, or governance infrastructure. We therefore interpret Control Assimilation not as the disappearance of the human from the loop, but as the incorporation of the route to human control into the technical architecture itself. Put differently, the human can remain formally authoritative while becoming operationally downstream on the runtime control path. This is a structural interpretation of the sampled architectures, not evidence that we observed a temporal progression within any individual system.

4.2 From out-of-the-loop performance to machine-mediated access

Classic automation research anticipated part of this problem. Bainbridge described the paradox of assigning humans responsibility for rare abnormal conditions after automation has removed them from routine control (1983), and Endsley and Kiris showed that passive supervisory roles can impair situation awareness and takeover performance (1995). Parasuraman and colleagues emphasized that automation can be distributed across information acquisition, analysis, decision selection, and action implementation rather than represented by a single scalar level (2000). Our findings extend this tradition from human cognitive state to control-path architecture. In the systems we reconstructed, attention was not only a scarce human cognitive resource; it was also an allocation variable controlled by routing, risk classification, confidence thresholds, escalation rules, failure detectors, or relevance filters. The key question is therefore not only whether the operator is cognitively 'in the loop', but whether the system architecture gives the operator a timely and sufficiently independent path into the loop when it matters.

4.3 Why 'human-in-the-loop' is not a control claim

The findings also clarify a limitation of binary HITL terminology. Presence is a topological fact; control is a causal capacity. A person can appear in an approval, monitoring, or appeal step without receiving an adequate intervention opportunity, without possessing an executable mechanism, or without being able to reverse the trajectory already created by upstream automation. Meaningful Human Control frameworks explicitly require alignment between responsibility, ability, authority, and reason-responsive system design (Santoni de Sio & van den Hoven, 2018; Cavalcante Siebert et al., 2023). Langer et al. add an outcome-side requirement: reliable detection of relevant errors is crucial for effective oversight, not merely the existence of a monitoring role (2025). Our architecture-level coding operationalized a complementary distinction by separating intervention opportunity, intervention capability, formal authority, and observed effective control. Their non-equivalence was empirically consequential: only 5/14 outcome-evaluable architectures contained direct human control exercise under the base attribution rule, and only 3/14 under the strict high-confidence rule, despite widespread documentation of approval gates, vetoes, overrides, and governance mechanisms. A HITL label should therefore not be interpreted as evidence that human control was exercised, effective, or even timely.

4.4 Three recurring control configurations locate the human boundary differently

The RC taxonomy is most useful as a map of where the human boundary sits. RC-A, the most frequent configuration within the frozen cohort (11/17), delegated routine execution while reserving selected risk, exception, promotion, or policy boundaries for human authorization. This is not continuous supervision; it is selective governance. RC-B (4/17) relied primarily on exception, transfer, monitoring, or takeover pathways. In these systems, the human may possess takeover authority but still depend on the system to detect the failure and recruit the operator, making takeover opportunity analytically distinct from takeover capability. RC-C (2/17) preserved final end-user acceptance, rejection, appeal, or physical action after

upstream machine filtering. Here, terminal human choice remained real, but it was not unmediated: the machine had already influenced salience, option exposure, timing, or framing. The gate-first sensitivity analysis deliberately collapsed RC-C into RC-A, showing that RC-C is a role-sensitive boundary configuration rather than a natural latent class. This sensitivity is a feature of the construct definition and is therefore made explicit rather than concealed. The independent blinded reclassification strengthens the claim that this locked decision rule is reproducible at the label level: all 17 cases were assigned to the same RC by the second reviewer (kappa = 1.00). This should not be overinterpreted as external validation of natural classes. The reviewer classified standardized neutral architecture summaries prepared from the frozen evidence base rather than independently re-extracting the primary sources, so the result supports rule reproducibility, not independent evidence extraction or taxonomy derivation.

4.5 Formal authority and effective control can diverge

The separation between governance architecture and observed effective control is central. Independent constraints, approval tokens, policy envelopes, escalation rules, and rollback mechanisms are substantively important; they should not be dismissed as symbolic merely because direct human action was not observed. However, their presence does not by itself demonstrate that a human detected the relevant event, understood it, intervened in time, or causally changed the outcome. This distinction matters because highly autonomous systems can preserve a chain of formal accountability even while the human's practical control surface is narrow. The resulting asymmetry resembles the responsibility-gap concern identified by Matthias (2004) and the false-assurance problem identified by Green (2022), but our evidence suggests a specific architectural mechanism: responsibility may remain human while the system conditions the informational and temporal circumstances under which that responsibility can be meaningfully exercised.

4.6 Implications for regulation and technical assurance

This interpretation has direct implications for oversight requirements. Article 14 of the EU AI Act does not merely require a named human overseer; it requires capabilities to understand, monitor, disregard, override, reverse, intervene, and stop high-risk systems (European Parliament & Council of the European Union, 2024). Our findings suggest that these requirements should be tested as end-to-end control paths rather than as interface or documentation checkboxes. A credible oversight audit should ask at least five questions: what mechanism recruits the human; what transformations occur before the human sees the case; what residual time and information remain at that point; which actions are technically executable and independent of the acting agent; and whether completed human review or intervention is actually observed. The failure of all 17 architectures to support an exact same-universe human-review proportion illustrates why routing, notification, display, and approval requirements cannot be substituted for evidence of completed review. Similarly, our ICC coding indicates that independent constraints must be evaluated at the enforcement boundary itself, not inferred from the existence of a policy statement.

Existing assurance and certification frameworks already operationalize important components of responsible human oversight. For example, the VDE VCIO trustworthiness framework includes explicit 'Responsible Human Oversight' indicators covering operator expertise, effort required to understand and

interact with the system, documented human control and involvement, rapid human takeover, and transparency or post-hoc change in human-on-the-loop settings (VDE, 2022). These requirements are complementary to, rather than displaced by, Control Assimilation. The additional assurance question raised by our findings is whether the upstream control path that supplies the human with cases, representations, timing, and intervention opportunities is itself examined. A capable human overseer or well-designed takeover interface cannot by itself establish meaningful operational access to control if the architecture that detects, filters, classifies, or routes control-relevant states remains unexamined.

As a design implication rather than an empirical effect estimate, the three recurring configurations suggest different assurance priorities. In RC-A, selective gate-based delegation, assurance should test the gate itself: the criteria that trigger human review, false-negative pathways, and independently sampled cases from the non-escalated routine stream. In RC-B, exception/takeover control, assurance should stress detection-to-recruitment latency, operator situation awareness at handover, and the residual intervention margin under failure or abnormal conditions. In RC-C, final human action after upstream filtering, assurance should examine how options or evidence were filtered, ranked, framed, or suppressed before terminal choice, including whether omitted alternatives remain contestable. Across all three configurations, meaningful oversight may therefore require controls to begin upstream of the final human interface and to be distributed across the technical and organizational supply chain, because a downstream operator cannot independently validate every prior detection, classification, representation, or routing decision.

4.7 Relationship to emerging agentic-AI oversight literature

The strongest conceptual overlap spans several recent lines of work. Nguyen and colleagues' 2026 survey of agentic-AI controllability distinguishes causal from normative control and organizes mechanisms into constraints/guardrails, adaptive control, agent-in-the-loop, and human-in-the-loop paradigms (2026). For AI systems based on foundation models, Tsamados, Floridi, and Taddeo argue that human control is better understood through human-machine teaming and collaborative agency than through exclusive human supervision (2025). Takahashi is especially close on the architecture of oversight access: humans and automated verifiers are modeled as costly interrupt channels characterized by routing error, delay, availability, congestion, selective review, and irreversibility constraints (2026). That prior work makes it untenable to claim novelty for the general propositions that HITL is insufficient, that routing and oversight latency matter, or that human review is a scarce channel. Our contribution is narrower but empirically distinct: we use the responsible human's actual control path as the cross-system unit of analysis and reconstruct, across concrete real/deployment-proximal architectures, the transformations that select, compress, classify, route, or time human intervention; the authority and executable capability that remain at the human boundary; independently enforceable constraints; and whether direct human control was actually exercised. Zhu and colleagues likewise distinguish AI operative agency from human evaluative agency (2026); Kumar and Singh treat human intervention as a dynamically allocated resource (2026); Dhanorkar and colleagues identify a priori control, co-planning, real-time monitoring, and post hoc review in developers' oversight work (2026); and Mitchell, Ghosh, and Passi argue that agent designs can erode capacities on which oversight depends (2026). Control Assimilation adds a cross-architecture empirical

proposition across these literatures: formal human authority can persist even when the informational, temporal, and technical conditions for exercising that authority are incorporated into the system architecture. Adjacent standards, certification schemes, and enterprise best-practice documents provide a complementary practical literature because they often encode operational oversight requirements without making the control path itself the unit of analysis; the VDE VCIO framework is one example (VDE, 2022). A structured comparison with the closest competing constructs is provided in Supplementary Table S6.

A further adjacent 2026 theory is Estep's 'instrumental succession,' which proposes that human controllers may progressively pursue instrumental dispositions on behalf of AI systems, allowing AI to exercise an increasing share of oversight and control and potentially shifting the locus of agency from humans to AI (2026). This is close to Control Assimilation in possible trajectory but distinct in analytic scale and evidentiary claim. Instrumental succession is explicitly longitudinal and concerns gradual transfer of agency; Control Assimilation, as operationalized here, is a cross-sectional structural condition that can be present while formal authority still remains human. It therefore does not require succession of agency, nor do our data establish that assimilation progresses monotonically over time.

Evidence credibility materially constrained the strength of this interpretation. In a sensitivity analysis restricted to the eight cases represented by a formal journal or conference publication channel, the architecture distribution was RC-A 5, RC-B 3, and RC-C 0. Local veto/override evidence exceeded documented global stop in all 8/8 cases, and clear system/policy-mediated attention recruitment remained present in 6/8. Among the seven outcome-evaluable cases in this restricted set, direct human control exercise was documented in only 1/7 under both the base and strict attribution rules. Thus, the direction of the local-control and mediated-access findings persisted, while RC-C disappeared and direct-exercise evidence became thinner (Supplementary Table S5). We therefore treat RC-C as a useful boundary configuration rather than a prevalence-stable empirical class, and we do not use publication status as a proxy for validity. Technical reports can contain unusually precise architecture evidence, while formal publication does not eliminate developer self-report or weak outcome attribution.

4.8 Design implication: oversight should be engineered as a control system

The practical implication is that human oversight should be engineered and evaluated as a control system, not treated as a role label. Systems should expose the provenance of escalation decisions, distinguish machine-selected attention from independently sampled review, test intervention latency under failure conditions, preserve independent stop or rollback boundaries where warranted, and report denominator-aware evidence of completed human review. Human-AI interaction work already shows that explanation alone does not reliably produce appropriate reliance and that forcing functions can reduce overreliance at a usability cost (Bucinca et al., 2021; Mehrotra et al., 2024). The analogous architecture lesson is that authority without a usable, timely, and independently enforceable pathway is insufficient evidence of control. Future evaluation should therefore measure not only system performance but also the integrity of the human control path. A useful external-validation stratum should deliberately include

standards, certification schemes, and enterprise best-practice documents alongside formal publications, because these sources may encode operational oversight designs that are implemented in practice but incompletely represented in concept-focused bibliographic literatures.

5. Limitations

This study has important limitations. First, the 17 architectures constitute a frozen first-wave analytic sample, not a probability sample of deployed agentic AI; observed RC proportions must not be interpreted as population prevalence. Second, although all 6,754 canonical records were routed prospectively, only the P1 first-wave lane (n = 1,156) was advanced through the intensive full-text workflow before the analysis freeze. The 5,598 P2 records remained indeterminate_requires_FT and were not substantively excluded. Completing P2 could identify additional architectures and alter the descriptive distribution; accordingly, this work should not be represented as exhaustive full-text adjudication of the complete search universe. Third, within P1, 166 records had not undergone substantive full-text adjudication at the freeze. Their non-adjudication status is reported transparently; the absence of a priority-stratum concentration does not establish random missingness. Fourth, evidence was heterogeneous in publication form, implementation maturity, and developer/self-report dependence. The formal-publication sensitivity retained only eight cases and eliminated RC-C, underscoring that some configuration patterns are evidence-source sensitive (Supplementary Table S5). Fifth, independent reclassification demonstrated perfect RC-label agreement on standardized neutral case summaries, but it was not an independent primary-source extraction: the summaries themselves were produced from the frozen primary analytic workflow. The result therefore supports reproducibility of the locked classification rule, not external validation of the underlying evidence extraction or independent derivation of the taxonomy. Sixth, RC-B was materially confounded with D2/embodied deployment, and the embodied stratum was too small for comparative inference. Seventh, A003 and A012 retained MODERATE primary configuration confidence, while RC-C depended on an explicit end-user/final-action precedence rule; the independent reviewer rated A012 even more cautiously (LOW) without changing its RC-B classification. Eighth, because empirical eligibility required a reconstructable human-machine control path, the universal presence of a positive MMCD path within the final architecture cohort is partly selection-dependent and must not be read as a prevalence estimate. Ninth, no architecture supported an exact same-universe human-review proportion, preventing review-rate comparison. Tenth, the human-control evidence attribution categories distinguish direct exercise from architecture-level control effects but do not estimate causal human effectiveness. Finally, Control Assimilation is introduced here as a descriptive architecture-level construct. Its predictive validity, threshold structure, relationship to failure risk, and generalizability require prospective testing, independent primary-source replication, and external validation in additional architecture samples.

6. Conclusions

Across this frozen cross-domain sample of highly autonomous and agentic architectures, human authority usually did not vanish. Instead, control was selective, local, and technically mediated. The human often retained approval, veto, override, appeal, or terminal-action authority while upstream system

or policy logic determined which cases were surfaced, how they were represented, and when intervention became possible. We define this architecture-level divergence as Control Assimilation. The construct does not imply that autonomy eliminates human agency, nor does the present cross-sectional evidence establish a temporal process of progressive assimilation; it identifies a structural mechanism by which formal authority can coexist with technically conditioned operational access to control.

For agentic AI, the critical safety question is therefore no longer merely 'Is a human in the loop?' It is: 'Who controls the human's access to the loop, and what remains executable when the human gets there?'

Declarations

Competing interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During preparation of this manuscript, the authors used OpenAI ChatGPT, including GPT-5.6 Sol and GPT-6 Astra (OpenAI), for structural and linguistic refinement of author-defined content. The authors reviewed and edited the outputs and take full responsibility for the content of the submitted work. AI-assisted research activities and their human verification are described in Section 2.10.

Funding

No specific funding was received for this work.

Data availability

The frozen data supporting this study are publicly available in Mendeley Data: Rustenbach, Christian; Salewski, Christoph; Kral, Annika; Reichert, Stefan (2026), “Frozen data underlying ‘Control Assimilation in Agentic AI’: architecture cases, human-control paths, screening status, and RC reproducibility”, Mendeley Data, Version 1, https://doi.org/10.17632/f8vbzyrry.1. Supplementary Tables S1–S6 report the frozen architecture-level results, source and manifestation provenance, full-text assessment status at the prespecified analysis freeze, case-level control-architecture coding, robustness and sensitivity analyses, evidence-credibility appraisal, and competing-concept differentiation. The deposited data preserve the first-wave analysis freeze; P2 records were not substantively adjudicated for this study.

References

1. Ahmed AG (2025) Develop an AI-Driven Fault Detection Model to Autonomously Troubleshoot Electrical Power Grids in High-Risk Offshore Oil Platforms. Br J Earth Sci Res 13(2):19–41. https://doi.org/10.37745/bjesr.2013/vol13n21941
2. Amershi S, Weld D, Vorvoreanu M et al (2019) Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233
3. Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779. https://doi.org/10.1016/0005-1098(83)90046-8
4. Bucinca Z, Malaya MB, Gajos KZ (2021) To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287
5. Cavalcante Siebert L, Lupetti ML, Aizenberg E et al (2023) Meaningful human control: actionable properties for AI system development. AI Ethics 3:241–255. https://doi.org/10.1007/s43681-022-00167-3
6. Chalupka R (2026) Machine-Native Intelligence v2: Operational Evidence from a 208-Daemon Autonomous System. RRC-AI Research Initiative. Zenodo. https://doi.org/10.5281/zenodo.19486388
7. Chittala S, Unnava N (2026) AI-Driven Release Agents: Leveraging AIOps and Generative AI Frameworks for Predictive and Reliable Software Delivery. J Int Crisis Risk Communication Res 9(1):35–44. https://doi.org/10.63278/jicrcr.vi.3565 Frozen evidence provenance additionally retains the linked Zenodo work family (Supplementary Table S1)
8. Dhanorkar S, Passi S, Vorvoreanu M (2026) Human oversight of agentic systems in practice: Examining the oversight work, challenges, and heuristics of developers using software agents. arXiv. https://doi.org/10.48550/arXiv.2606.05391
9. Endsley MR, Kiris EO (1995) The out-of-the-loop performance problem and level of control in automation. Hum Factors 37(2):381–394. https://doi.org/10.1518/001872095779064555
10. Estep PW (2026) Instrumental succession: the gradual transfer of agency to AI. Front Psychol 17:1922407. https://doi.org/10.3389/fpsyg.2026.1922407
11. European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14: Human oversight. Official Journal of the European Union
12. Green B (2022) The flaws of policies requiring human oversight of government algorithms. Comput Law Secur Rev 45:105681. https://doi.org/10.1016/j.clsr.2022.105681
13. Idris D (2026) Autonomous Defense Transformers: Security-Native Reasoning for Digital Infrastructure. Technical paper. Zenodo. https://doi.org/10.5281/zenodo.18878123
14. Kumar M, Singh D (2026) Balancing autonomy and oversight in reliable agentic artificial intelligence through adaptive human interaction architectures. Discover Artif Intell 6:709. https://doi.org/10.1007/s44163-026-01373-2
15. Langer M, Baum K, Schlicker N (2025) Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. Mind Mach 1.

35https://doi.org/10.1007/s11023-024-09701-0

16. Maheswaran R, Rogers CM, Sanchez R, Szekely PA, Gati G, Smyth K, VanBuskirk C (2010) Multi-agent systems for the real world. In 2010 International Symposium on Collaborative Technologies and Systems (CTS). IEEE. https://doi.org/10.1109/CTS.2010.5478452
17. Malik A, Jayasinghe D, Klemick B, Shah P, Talasu N, Trivedi VT (2026) Decentralized Granular Access Control for Agentic AI Systems in Critical Infrastructure. arXiv. https://doi.org/10.48550/arXiv.2607.22611
18. Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183. https://doi.org/10.1007/s10676-004-3422-1
19. Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML (2024) A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. ACM J Responsible Comput 1(4):26. https://doi.org/10.1145/3696449
20. Mitchell M, Ghosh A, Passi S (2026) AI Agents Push Humans Out of the Loop. arXiv. https://doi.org/10.48550/arXiv.2608.23642
21. Mukherjee S, Sanz Ausin M, Aggarwal K, Datta D, Puri S, Jin W et al (2026) Perfecting Human-AI Interaction at Clinical Scale: Turning Production Signals into Safer, More Human Conversations. *arXiv*. https://doi.org/10.48550/arXiv.2603.29893
22. Nguyen MH, Nguyen DH, O'Sullivan B, Nguyen HD (2026) On Controllability in Agentic AI: A Survey. Mind Mach 36:29. https://doi.org/10.1007/s11023-026-09783-y
23. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410. https://doi.org/10.1177/0018720810376055
24. Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybernetics - Part A: Syst Hum 30(3):286–297. https://doi.org/10.1109/3468.844354
25. Pratinav A, Kataria A (2026) AIDR-Cloud: An Agentic AI-Driven Incident Response Framework for Cloud Failures. In SoutheastCon 2026. IEEE. https://doi.org/10.1109/SOUTHEASTCON63549.2026.11476571
26. Ravi CS, Patlolla RR (2024) Autonomous Incident Remediation: A Closed-Loop RAG Framework for Real-Time Root Cause Analysis and Knowledge Synthesis in Distributed Cloud Systems. Int J Adv Res Sci Communication Technol. https://doi.org/10.48175/IJARSCT-19600A
27. Ravi Kumar SG (2026) Agentic Data Pipeline Monitoring: A Multi-Agent Architecture for Anomaly Detection in Financial Transaction Systems. Technical report. Zenodo. https://doi.org/10.5281/zenodo.20549210
28. de Santoni F, van den Hoven J (2018) Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015
29. Sharma A (2026) AEGIS: Evidence-Driven Runtime Governance for Autonomous AI Agents. ANKR Labs technical report. Zenodo. https://doi.org/10.5281/zenodo.19998776

30. Sweet KE Jr. (2026) CMPSBL OS Substrate v6.x.x - A Governed Cognitive Operating System for Persistent, Self-Improving AI Systems. Technical documentation. Zenodo. https://doi.org/10.5281/zenodo.18234909 Exact concrete version v6.x.x: https://doi.org/10.5281/zenodo.18425208
31. Takahashi K (2026) Oversight-Centered Metrology and Control for Agentic Systems: Costly Interrupt Channels, Claim Margins, and Deployment-Relevant Evaluation. Zenodo. https://doi.org/10.5281/zenodo.18973272
32. Teixeira RN (2026) Gamifying Sustainability: Harnessing AI and IoT for Human-Centric Energy Efficiency [Master's thesis, Faculty of Engineering, University of Porto]
33. Tsamados A, Floridi L, Taddeo M (2025) Human control of AI systems: from supervision to teaming. AI Ethics 5:1535–1548. https://doi.org/10.1007/s43681-024-00489-4
34. Vangoor VKR (2020) *Autonomous Infrastructure Provisioning Using AI-Driven DevOps Automation Framework*. Zenodo. https://doi.org/10.5281/zenodo.19183280
35. VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V (2022) VCIO based description of systems for AI trustworthiness characterisation: VDE SPEC 90012 V1.0 (en). VDE
36. Vinjamuri VRVV (2026) Human-AI Collaborative Frameworks for Intelligent Fault Detection, Proactive Quality Management, and Augmented Decision Support in Fixed Wireless Access Networks. J Inform Syst Eng Manage 11(3s). https://doi.org/10.52783/jisem.v11i3s.15183
37. Yu W, Lv J, Ying Z, Jin Y, Wen C, Lu C (2026) ARMADA: Autonomous Online Failure Detection and Human Shared Control Empower Scalable Real-World Deployment and Adaptation. IEEE Rob Autom Lett. https://doi.org/10.1109/LRA.2026.3709635
38. Zhang C, Xu Z, Liu J, Li D, Gu H, Zhang P, Gu N, Lu T (2026) Transparent and Controllable Recommendation Filtering via Multimodal Multi-Agent Collaboration. arXiv. https://doi.org/10.48550/arXiv.2604.17459
39. Zhu L, Lu Q, Ding M, Lee SU, Wang C (2026) Designing meaningful human oversight in AI. AI Ethics 6:286. https://doi.org/10.1007/s43681-026-01147-7

Tables

Table 1
Frozen architecture cases and recurring control configurations.

Case	System / architecture	Domain	Stratum / deployment	RC	Outcome evaluable	Primary reference
A001	ARMADA multi-robot shared control	Robotics/shared control	Embodied / D2	RC-B	Yes	Yu et al. (2026)
A002	AIDR-Cloud	Cloud incident response	Digital / D1	RC-A	Yes	Pratinav & Kataria (2026)
A003	CTS real-world coordinator MAS	Field multi-agent coordination	Digital / D2	RC-B*	Yes	Maheswaran et al. (2010)
A004	Polaris clinical safety constellation	Clinical conversational AI	Digital / D1	RC-B	Yes	Mukherjee et al. (2026)
A005	AEGIS runtime governance	Agent runtime governance	Digital / D1	RC-A	Yes	Sharma (2026)
A006	Delegated identity/access control	Critical-cloud IAM	Digital / D1	RC-A	Yes	Malik et al. (2026)
A007	Tiered FWA quality management	Telecommunications	Digital / D1	RC-A	Yes	Vinjamuri (2026)
A008	Machine-Native Intelligence v2	Autonomous software	Digital / D1	RC-A	Yes	Chalupka (2026)
A009	Controllable recommendation filtering	Recommender systems	Digital / D1	RC-C	Yes	Zhang et al. (2026)
A010	ADAF provisioning/self-healing	Multi-cloud DevOps	Digital / D1	RC-A	Yes	Vangoor (2020)
A011	Closed-loop cloud remediation	Cloud remediation	Digital / D1	RC-A	Yes	Ravi & Patlolla (2024)
A012	Offshore power-grid fault management	Critical infrastructure	Embodied / D1	RC-B*	No	Ahmed (2025)
A013	AI-Driven Release Agents	Software delivery	Digital / D1	RC-A	Yes	Chittala & Unnava (2026)
A014	CMPSBL OS v6.x.x	Cognitive operating system	Digital / D1	RC-A	No	Sweet Jr. (2026)

Case	System / architecture	Domain	Stratum / deployment	RC	Outcome evaluable	Primary reference
A015	Agentic financial pipeline monitoring	Financial data infrastructure	Digital / D1	RC-A	Yes	Ravi Kumar (2026)
A016	RL/IoT energy intervention	Human-centric energy/IoT	Digital / D1	RC-C	Yes	Teixeira (2026)
A017	PulseADT autonomous defense	Cybersecurity	Digital / D1	RC-A	No	Idris (2026)

**RC-B membership MODERATE for A003 and A012; all other RC memberships HIGH. Complete source/manifestation provenance is in Supplementary Table S1; case-level coding is in Supplementary Table S3.*

Table 2
Human-control evidence attribution categories among outcome-evaluable architecture cases (n = 14).

Evidence category	n	%	Cases	Interpretive meaning
High-confidence direct human exercise	3	21.4%	A001, A009, A016	Actual human takeover/appeal/final action explicitly exercised with sufficiently direct attribution.
Moderate-confidence direct human exercise	2	14.3%	A008, A015	Human act explicit, but narrow/test-specific or architecture-level attribution limited.
Control effect observed; human contribution not isolated	7	50.0%	A002, A004-A007, A010-A011	Operational gate/block/escalation consequence observed without separable human contribution.
Indirect / primarily machine-control relevance	2	14.3%	A003, A013	Outcome relevant to architecture/recoverability but not a narrow exercised human-control consequence.

Base direct-human coding = 5/14 (35.7%); strict HIGH-confidence direct coding = 3/14 (21.4%). These are evidence-type counts, not effectiveness rates.

Figures

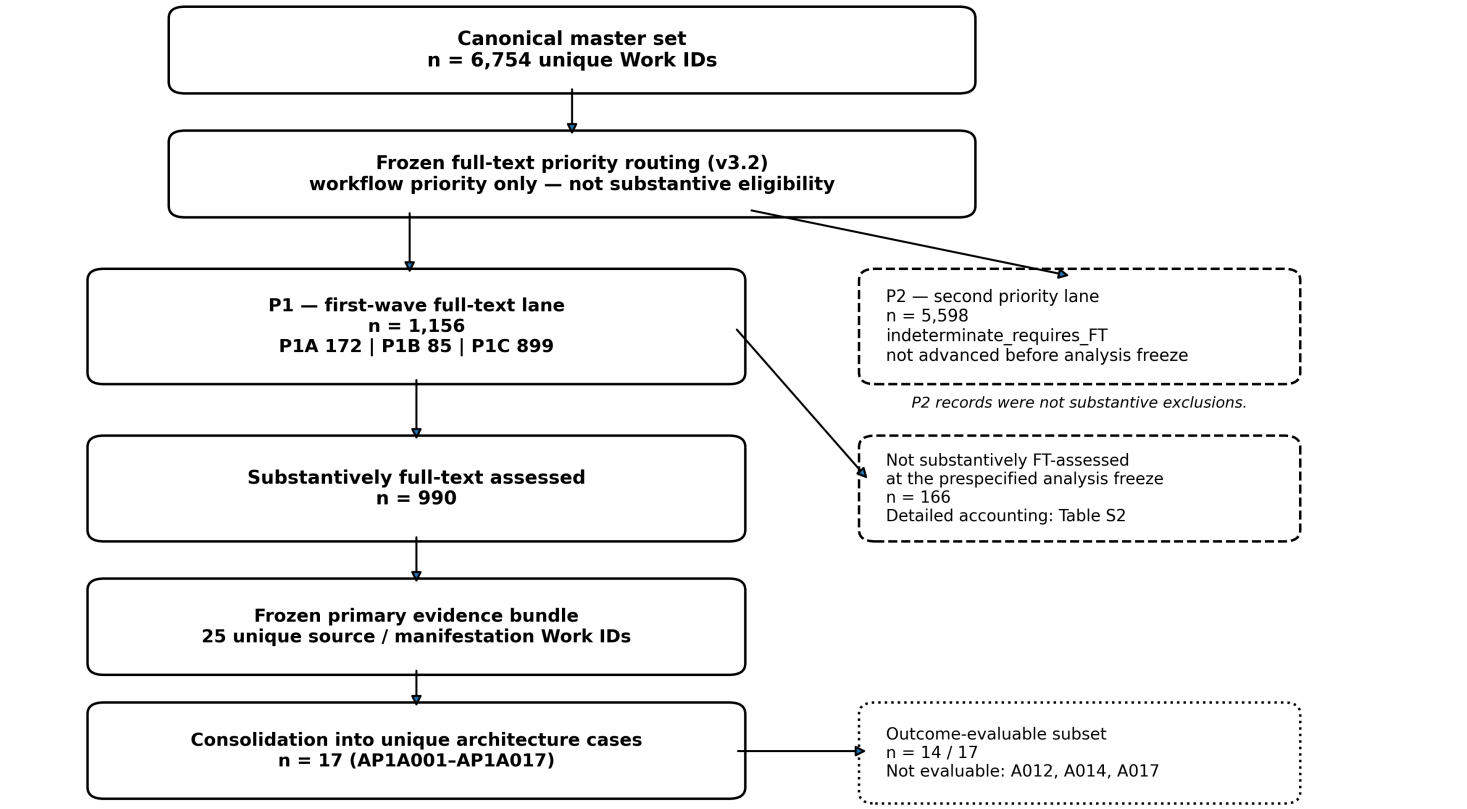


Figure 1

Study identification, frozen priority routing, full-text assessment, and architecture-case consolidation.

All 6,754 canonical Work IDs were assigned to the frozen priority router. P1/P2 ordering was a workflow device rather than a substantive eligibility decision; P2 remained indeterminate_requires_FT and was not advanced before the analysis freeze. The 166 P1 records without substantive full-text adjudication are detailed in Supplementary Table S2.

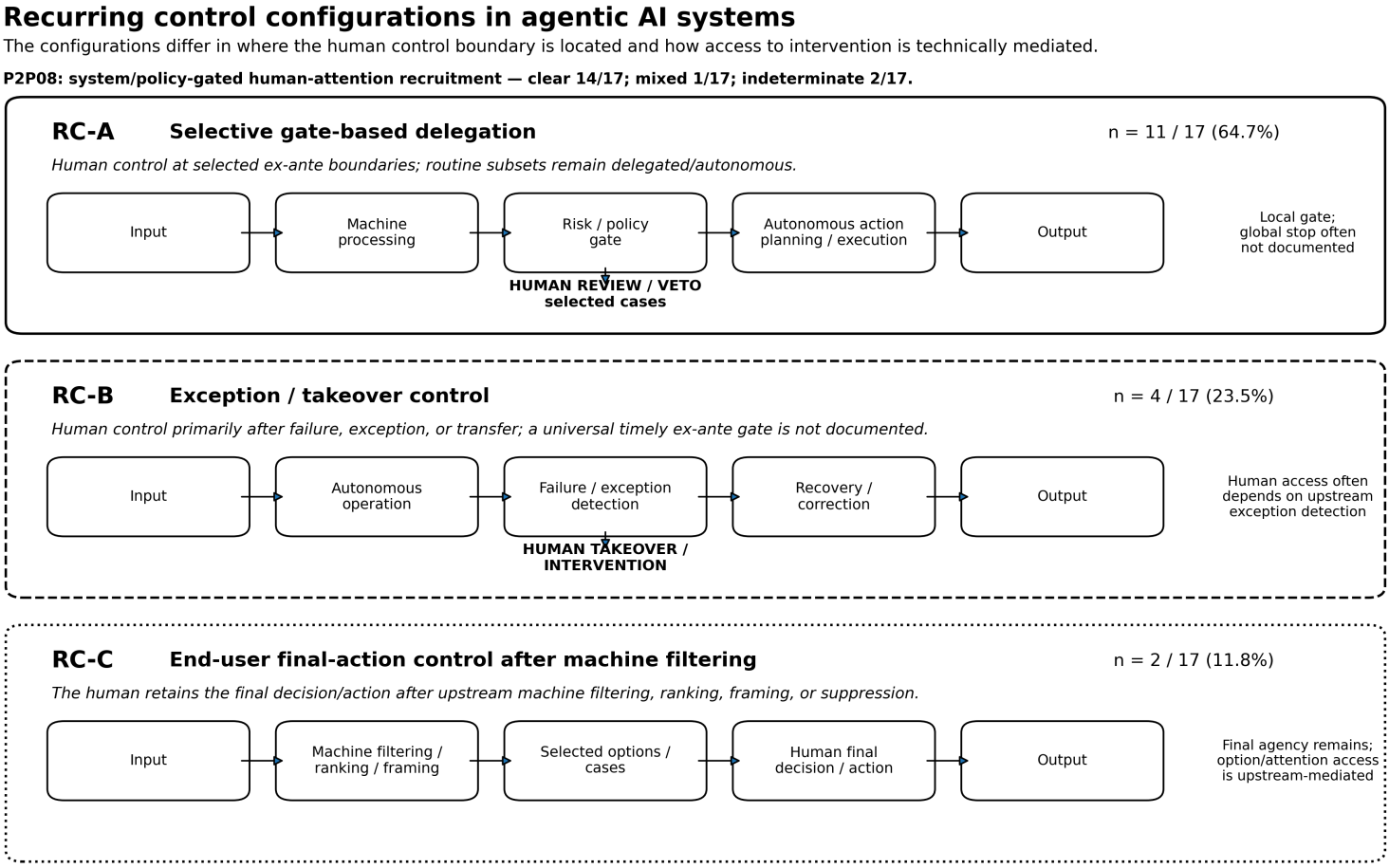


Figure 2

Recurring control configurations in agentic AI systems.

RC-A: selective gate-based delegation. RC-B: exception/takeover control without a clear universal timely pre-action gate. RC-C: end-user final-action/appeal control after upstream machine filtering. The taxonomy is descriptive and non-causal.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CAControlAssimilationPREPRINTSupplementv1.120260925.pdf