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.

From Human Oversight to Effective Control: A Socio-Technical Safety-Control Framework for High-Risk AI Systems

Karim Hardy

hardykarim@gmail.com

Embry–Riddle Aeronautical University

Research Article

Keywords: Human oversight, Meaningful human control, High-risk AI, Human–AI interaction, Socio-technical systems

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10151972/v1

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

Additional Declarations: No competing interests reported.

Abstract

Human oversight is widely prescribed as a safeguard for high-risk artificial intelligence, yet the presence of a human reviewer does not by itself establish effective control. This study asks under what socio-technical conditions oversight can detect, interpret, challenge, and alter an AI-mediated decision trajectory rather than operate as a nominal procedural requirement. It develops an Oversight-to-Control (O2C) framework comprising six functions—Observe, Detect, Interpret, Challenge and Decide, Intervene, and Recover and Learn—supported by competence, observability, time, authority, actionability, escalation, logging, and organizational learning. The framework is evaluated against a structured qualitative documentary corpus containing eight regulatory and risk-management instruments, nineteen high-impact AI or automated-decision cases, and forty-one verified Oversight-Control Evidence Statements. Cross-framework coverage analysis, cross-case comparison, mechanism coding, negative-case analysis, triangulation, and seven sensitivity analyses were applied. The prescriptive instruments converge most strongly on challenge and intervention, while time and workload are less consistently specified. In the selected cases, interpretation was degraded in all eighteen assessable cases and timely intervention in seventeen of nineteen. Missing escalation or fallback, inadequate validation or monitoring, and organizational governance failure were the leading coded mechanisms. Recovery and learning were often documented only after delayed or failed intervention. The findings support O2C as an analytically developed diagnostic framework that distinguishes nominal oversight, partial control, workflow fragility, and multi-level control breakdown. The framework is benchmarked against public evidence rather than externally validated, and its counts do not estimate population prevalence.

1 Introduction

Artificial intelligence increasingly mediates decisions that shape access to employment, education, credit, health care, public benefits, mobility, and public authority. In these settings, the consequences of an erroneous, discriminatory, or poorly governed output are not confined to model performance. They may alter a person's legal position, material resources, safety, or ability to contest an institutional decision. Human oversight has therefore become a central safeguard in policy and governance frameworks for high-impact and legally regulated AI. Yet sociotechnical accounts of AI caution against treating "the human" and "the system" as isolated actors. Decisions emerge from configurations of interfaces, data, organizational rules, professional roles, institutional incentives, and regulatory structures (Steen et al. 2025; Shneiderman 2020). Accordingly, the value of oversight depends less on whether a human appears somewhere in a workflow than on whether that person and the surrounding organization can materially influence what the system does.

A nominally human-in-the-loop arrangement can preserve the appearance of control while withholding its practical conditions. A reviewer may see an output but lack the contextual information needed to recognize an anomaly. The output may be technically explained yet not intelligible for the decision at hand. Time pressure, alert volume, or routinized workflow may prevent an independent assessment. The reviewer may understand that an output is problematic but lack formal authority, an actionable interface, or an escalation route. Alternatively, a challenge may be possible only through an appeal after the decision has already caused harm. These configurations place responsibility on human actors without granting commensurate control. Green (2022) argues that mandatory human oversight policies can therefore become ineffective or even counterproductive when institutions do not supply the capabilities and incentives required for meaningful review. Robbins (2025) similarly separates the prior question of what should be delegated to machines from the design of safeguards after delegation. Both perspectives imply that human presence is not the appropriate unit of evaluation; practical control capacity is.

Several research traditions illuminate parts of this problem. Work on meaningful human control examines whether automated systems track human reasons and whether responsibility can be traced to informed human agents (Santoni de Sio and van den Hoven 2018). Human-factors research identifies automation bias, selective adherence, trust calibration, workload, and interaction design as influences on human–AI performance (Alon-Barkat and Busuioc 2023; Romeo and Conti 2026). Explainable-AI research asks whether users can understand model behavior and evaluate outputs, while accountability and audit scholarship focuses on governance across the system lifecycle (Miller 2019; Raji et al. 2020). In parallel, analyses of AI ethics guidelines show substantial convergence around principles such as transparency, human agency, safety, and accountability, alongside persistent gaps in enforcement and operationalization (Jobin et al. 2019; Mittelstadt 2019; Morley et al. 2020;

Hagendorff 2020). These literatures establish that oversight is not a single act. However, they do not yet provide a common operational sequence that connects the information available to a reviewer, the reviewer's independent judgment, the authority to intervene, and the organization's ability to recover and learn.

Recent contributions in AI & SOCIETY sharpen the institutional and human dimensions of the debate. Laux (2024) conceptualizes oversight as institutionalized distrust and emphasizes the democratic design of governance under the European Union AI Act. Te'eni et al. (2026) treat control by design as a feedback process sustained by human learning. Romeo and Conti (2026) synthesize automation-bias research and its implications for explainable AI, while Steen et al. (2025) argue for sociotechnical and cybernetic images that represent AI relations more adequately. These studies clarify why oversight must be designed rather than merely declared. The present study differs in its evidentiary and analytical combination. It compares the coverage of oversight functions across multiple regulatory and risk-management frameworks, examines how those functions break down across a multi-domain set of public cases, and links the two layers through a requirement–capability–evidence matrix. The resulting contribution is neither a purely normative theory nor a prevalence study. It is a structured documentary analysis of when oversight arrangements can plausibly perform a safety-control function.

The study addresses the following question: under what socio-technical conditions does human oversight in high-risk AI systems operate as an effective safety-control function rather than as a nominal procedural safeguard? Three subquestions organize the analysis. First, how do leading regulatory and risk-management frameworks specify the functions, resources, authority, and intervention mechanisms required for oversight? Second, which oversight-capability breakdowns recur in publicly documented high-impact AI incidents, and at what point in the oversight-to-control process do they occur? Third, which gaps between prescribed requirements and observed capabilities should an operational assessment framework expose? The expression "high-risk AI" is used in the article title to identify the target governance problem. Within the case analysis, however, high impact is kept separate from the formal legal classification of a high-risk system under the EU AI Act. Cases outside the European Union and cases predating the Act are treated as analytical analogues rather than assigned an unsupported legal status.

The article makes three related contributions. Its principal contribution is the Oversight-to-Control (O2C) framework, which models oversight through six functions: Observe, Detect, Interpret, Challenge and Decide, Intervene, and Recover and Learn. The functions are enabled by cross-cutting conditions including competence, observability, interpretability, time and workload, authority and independence, interface actionability, override and stop mechanisms, escalation, logging, feedback, and organizational learning. The second contribution is a mechanism-based typology of oversight failures that avoids the terminal explanation of "human error." The third is a requirement–capability–evidence matrix that connects what governance frameworks prescribe to what public cases demonstrate, limit, or leave uncertain. The analysis is based on eight prescriptive documents, nineteen cases, and forty-one verified evidence statements, with robustness assessed through rule-stability checks, triangulation, negative cases, and sensitivity analyses. The next section develops the conceptual foundations of O2C, followed by methods, results, discussion, limitations, and conclusion.

2 Human Oversight as Socio-Technical Control

2.1 Meaningful human control and delegation

Meaningful human control provides a foundational vocabulary for asking whether human agency remains substantively connected to automated behavior. Santoni de Sio and van den Hoven (2018) distinguish two conditions. Tracking concerns whether a system responds to the relevant human reasons for action; tracing concerns whether informed human agents can understand their role and be held responsible. Control can be meaningful even when an automated system performs complex functions, provided that its behavior remains appropriately connected to human reasons and responsibility. Conversely, a system can include a formal approval step while severing that connection. For oversight, the reviewer's position must be assessed in relation to the decision process, available information, and capacity to alter subsequent action. O2C takes this relational view as its starting point but specifies the operational functions through which tracking and tracing can be enacted.

Later work has translated meaningful control into more actionable design properties. Siebert et al. (2023) identify properties intended to make control implementable in AI system design, including the allocation of authority that matches responsibility. Yet they remain context-sensitive. Boshuijzen-van Burken et al. (2025) show that the value assigned to human control varies with system purpose, institutional setting, and competing design values. Te'eni et al. (2026) add that control by design depends on feedback and human learning over time. These accounts suggest that no single architectural label—human-in-the-loop, human-on-the-loop, or human-in-command—can establish control independently of context. Authority that is adequate in a slow administrative process may be ineffective in a safety-critical workflow, while expertise may be neutralized by opacity or a non-actionable interface. O2C therefore treats control as a configuration of functions and enabling conditions rather than a fixed level of human involvement.

The question of oversight also follows a prior decision about delegation. Robbins (2025) argues that some tasks should not be delegated to machines because doing so conflicts with values or responsibilities that cannot be preserved by downstream safeguards. Where delegation is permitted, however, the existence of a human review requirement still does not guarantee control. Green (2022) shows why mandatory human oversight can fail when reviewers lack expertise, incentives, institutional support, or an effective way to depart from the algorithm. Laux (2024) similarly frames oversight as an institutional design problem: distrust must be organized through roles, competencies, and powers rather than assumed to arise from an individual's conscientiousness. These critiques justify a capability-based test. Oversight should be considered meaningful only when the arrangement can support independent judgment and alter the decision trajectory within the relevant time window.

2.2 Human factors, reliance and interpretability

Human-factors research explains why formal access to an output may not lead to independent evaluation. Automation bias describes a tendency to over-rely on automated recommendations, including omission errors when a person fails to act because the system provides no alert and commission errors when incorrect advice is followed. Romeo and Conti (2026) show that the phenomenon remains important in human–AI collaboration, but its occurrence depends on task design, explanation, expertise, and organizational context. In public-sector experiments, Alon-Barkat and Busuioc (2023) identify both automation bias and selective adherence: decision-makers may follow algorithmic advice unevenly rather than uniformly. This distinction is important for oversight analysis. Agreement with the system is not proof of rubber-stamping, and disagreement is not proof of effective control. Evidence is needed about the quality and independence of the review, the information available, and the reviewer's ability to act. O2C therefore treats automation bias as one mechanism—F04—within a broader typology that also includes missing escalation, poor validation, weak reporting, and organizational governance failure.

The broader concept of trust likewise requires careful specification. Jacovi et al. (2021) distinguish warranted from unwarranted trust by considering the user's vulnerability and the implicit contract the system is expected to satisfy. Zhang et al. (2020) find that confidence information can affect trust calibration, yet better calibration does not necessarily produce better joint decisions. Schemmer et al. (2023) operationalize appropriate reliance as accepting correct AI advice while rejecting incorrect advice, which is more demanding than measuring general trust or agreement. These studies imply that oversight should not seek maximum trust or maximum skepticism. It should support a calibrated capacity to rely, challenge, and intervene. That capacity is partly cognitive, but it also depends on governance. A reviewer cannot appropriately reject an output when deviations are penalized or no practical override exists. Trust and reliance therefore belong inside a larger control structure rather than functioning as stand-alone psychological variables.

Interpretability is one condition of that structure, but it is not a universal solution. Miller (2019) shows that explanations are selective, contrastive, and socially situated; an explanation that is technically complete may be irrelevant to the decision-maker's actual question. Murdoch et al. (2019) likewise define interpretability in relation to predictive accuracy, descriptive accuracy, and relevance for an audience and purpose. Empirical studies complicate the common assumption that transparency necessarily improves control. Hase and Bansal (2020) find that many explanation methods do not reliably help users predict model behavior, and Poursabzi-Sangdeh et al. (2018) report that increased transparency can fail to improve error detection. The relevant O2C function is therefore not the production of an explanation as such, but the reviewer's capacity to interpret the output sufficiently to assess reliability and consequences. This definition allows the analysis to distinguish model opacity, task-level interpretability, and the practical ability to challenge a decision.

Interaction design can strengthen or weaken this capacity. Amershi et al. (2019) propose guidelines covering expectation setting, feedback, correction, and recovery across the human–AI interaction lifecycle. Buçinca et al. (2021) show that cognitive forcing functions can reduce over-reliance, although they may impose effort and performance costs. Bansal et al. (2021) find that explanations can increase acceptance without necessarily improving complementary team performance. Vasconcelos et al. (2023) further demonstrate that engagement with explanations depends on their perceived costs and benefits, and that explanations can reduce over-reliance in some conditions. Effective oversight requires interaction features that are aligned with task stakes, reviewer expertise, time, and available interventions. Interfaces may improve understanding while leaving authority unchanged, and overrides may remain unusable under time pressure. These dependencies motivate the separation of Interpret, Challenge and Decide, and Intervene within O2C.

2.3 Sociotechnical safety-control

A sociotechnical perspective expands the unit of analysis beyond the individual reviewer. Steen et al. (2025) argue that prevailing images of AI often misrepresent the relationships among people, technologies, and institutions. Shneiderman (2020) similarly presents human-centered AI as a design space in which high levels of automation can coexist with high levels of human control. Suresh and Guttag (2021) show that harms may enter at multiple stages of the machine-learning lifecycle, from data and problem formulation through deployment. Oversight at the final output stage cannot compensate automatically for upstream objective, data, validation, or organizational failures. The human reviewer is therefore situated within a control structure that allocates information, responsibilities, resources, and escalation pathways. Local interface, monitoring, and appeal problems may originate in upstream organizational decisions. O2C uses a constant functional vocabulary across these levels while preserving the possibility that one case involves several mechanism layers.

Feedback and learning are central to this structure. Te'eni et al. (2026) conceptualize control as an ongoing process in which human learning and system feedback affect future performance. Raji et al. (2020) place algorithmic auditing within an end-to-end lifecycle, connecting documentation, accountability, assessment, and corrective action. Li et al. (2021) likewise argue that trustworthy AI requires governance and monitoring practices that extend beyond abstract principles. These perspectives justify including Recover and Learn as a distinct O2C function. However, recovery must be temporally distinguished from timely control. An organization may investigate an incident, reverse a decision, or reform a process after substantial harm. Such learning can improve future control, but it does not show that the original oversight arrangement was effective. O2C therefore records recovery and organizational learning while separately assessing whether the human or institution could intervene before or during the harmful trajectory.

2.4 Governance, accountability and the principles-to-practice gap

Governance scholarship further demonstrates that oversight is an institutional rather than purely interpersonal relationship. Laux (2024) describes institutionalized distrust as a democratic design principle: systems should be organized so that claims and outputs can be questioned, reviewed, and constrained. Raji et al. (2020) similarly emphasize the infrastructure needed to produce evidence, assign responsibilities, and close accountability gaps through internal audit. In this view, oversight is not synonymous with accountability. Accountability concerns the obligation to explain, justify, and answer for decisions; oversight concerns the activities through which a system is monitored and challenged; effective control concerns the capacity to alter its course. An institution can assign accountability after an event without providing timely intervention authority. O2C keeps these distinctions explicit so that governance arrangements can be assessed by their operational effects.

The global AI ethics landscape reinforces the need for this distinction. Jobin et al. (2019) identify convergence around principles including transparency, justice, non-maleficence, responsibility, and privacy, but also variation in how these principles are interpreted and implemented. Mittelstadt (2019) argues that principles alone cannot guarantee ethical AI because they lack the institutional conditions and professional practices that make principles actionable. Morley et al. (2020) review tools and methods intended to translate "what" into "how," while Hagendorff (2020) highlights omissions related to enforcement and practical conflict resolution. Human oversight can become another high-level principle unless its functions, evidence, and powers are specified. A useful assessment must ask what the reviewer can observe, how anomalies are detected, what is interpretable, whether independent judgment is possible, what intervention mechanisms exist, and how the organization responds when control fails.

Lifecycle approaches offer a bridge from principles to practice. Li et al. (2021) connect trustworthy-AI objectives to governance, testing, monitoring, and implementation throughout system development and use. Suresh and Guttag (2021) map sources of harm across the lifecycle and thereby show why downstream review may be structurally incapable of correcting certain upstream choices. For oversight, this means that the relevant question is not only whether a human can inspect an output. It is whether the broader system supplies signals, records, responsibilities, and feedback loops that make detection and intervention possible. An integrated framework must diagnose where control is lost without treating every governance weakness as the same failure.

2.5 The O2C framework

The Oversight-to-Control framework addresses these requirements through six sequential but interacting functions. Observe concerns access to relevant system state, inputs, outputs, context, and known limitations. Detect concerns recognition of anomalies, limit violations, bias signals, or unsafe states. Interpret concerns understanding the significance, reliability, and possible consequences of an output. Challenge and Decide concerns independent judgment, contestation, rejection, or escalation. Intervene concerns the practical ability to modify, suspend, reverse, override, or stop the decision or action. Recover and Learn concerns containment, restoration, investigation, feedback, and organizational change. These functions are supported by competence, observability, interpretability, time and workload, authority, independence, incentives, interface actionability, override and stop mechanisms, escalation, a safe-state concept, logging, feedback, and organizational learning. Figure 1 summarizes the framework, and Table 2 provides the verified coding rules used in the analysis.

O2C does not assume that every case requires identical control arrangements or that the functions form a validated psychometric scale. It specifies a context-appropriate control chain that can be tested against documentary evidence. A breakdown early in the chain may make later authority unusable: a reviewer cannot intervene effectively in an anomaly that is not observable or interpretable. Conversely, accurate information does not establish control if the reviewer cannot challenge the system or act in time. Recovery and accountability may be robust even when operational intervention fails. It expects nominal oversight where responsibility lacks information or actionability, multi-level rather than purely human error, and gaps between prescribed and demonstrated capability. The following methods evaluate these expectations against the structured corpus.

3 Methods

3.1 Design and research questions

The study used a structured qualitative comparative documentary design to examine how human oversight is specified and how it breaks down in public evidence. Three analytical components were integrated. First, a directed content analysis compared oversight requirements across regulatory, intergovernmental, risk-management, and audit-guidance instruments. Second, a multiple-case documentary analysis examined publicly documented incidents, disputes, investigations, and deployment studies involving high-impact AI or automated decision systems. Third, the findings from the two layers were compared abductively to develop and refine the O2C framework. The design addressed RQ1 through framework-coverage analysis, RQ2 through mechanism coding and cross-case profiles, and RQ3 through a requirement–capability–evidence gap matrix. The design accommodates technical, cognitive, organizational, legal, and institutional evidence. As with end-to-end accountability and human–AI decision research, the analytical emphasis was on traceable processes, roles, and evidence rather than an isolated accuracy metric (Raji et al. 2020; Lai et al. 2021). The study was not designed to estimate causal effects or the prevalence of failure modes. Instead, it sought patterned correspondences among prescribed functions, documented capabilities, and control breakdowns. The comparative logic was therefore mechanism-oriented: a case contributed evidence when the record made it possible to identify an oversight actor, the point of human involvement, the information or authority available, and the resulting intervention or recovery pathway.

3.2 Corpus, evidence hierarchy and OCES

The frozen Phase 2 v1.0 corpus contained eight prescriptive documents and nineteen incident or deployment cases. The prescriptive layer included the EU Artificial Intelligence Act, the NIST AI Risk Management Framework and Playbook, the OECD AI Principles, the UNESCO Recommendation on the Ethics of Artificial Intelligence, the Council of Europe Framework Convention,

the UK Information Commissioner's Office human-review guidance, and the European Commission High-Level Expert Group's Ethics Guidelines for Trustworthy AI (European Commission HLEG 2019; European Union 2024; NIST 2023, n.d.; OECD 2024; UNESCO 2021; Council of Europe 2024; ICO n.d.). The Canadian Directive on Automated Decision-Making (P008) remained a contextual benchmark because a frozen current text was not independently captured, and ISO/IEC 42001 (P009) remained a public-overview benchmark because no licensed full text was available. The incident layer included nineteen cases across seven harmonized domain groups; sixteen relied on official judicial, regulatory, audit, investigation, or commission evidence and three on peer-reviewed deployment or validation studies. The DUO matter (C003) remained a discovery/sensitivity case because the complete decision and annexes were unavailable. Table 1 summarizes the corpus. The nineteen cases were harmonized from fourteen original domain labels into seven analytical groups: autonomous mobility (three), education (one), employment (two), financial services (one), health care (three), justice/policing/biometrics (three), and social protection/public administration (six). The event period began in 2011 to retain foundational automated-decision disputes and extended to 2023, while source versions and the final corpus were frozen on 24 June 2026.

Table 1
Characteristics of the prescriptive and incident/deployment corpora used in the analysis.

Category	Metric	Value	Denominator	Note
Prescriptive corpus	Documents included	8	8	P008 and P009 remain contextual only.
Incident corpus	Cases included	19	19	C003 remains discovery/sensitivity only.
Incident corpus	Official-primary cases	16	19	84.2% of included cases.
Incident corpus	Peer-reviewed B-tier cases	3	19	C017–C019.
Evidence statements	OCES included	41	41	22 prescriptive and 19 incident OCES.
Evidence statements	Normative requirements	22	41	Prescriptive layer.
Evidence statements	Failure or contrast records	19	41	Incident layer.
Domains	Original domain labels	14		Harmonized into seven analytical groups.
Domains	Analytical domain groups	7		Autonomous mobility: 3; Education: 1; Employment: 2; Financial services: 1; Healthcare: 3; Justice / policing / biometrics: 3; Social protection / public administration: 6
Profiles	Multi-function breakdown	5	19	At least four Failed statuses among the first five O2C functions.
Profiles	Upstream information / interpretation breakdown	7	19	Detect or Interpret failed without meeting the full-chain criterion.
Profiles	Authority / escalation and delayed intervention	3	19	Information was not fully failed, but action remained constrained or delayed.
Profiles	Partial control / model-workflow limitation	4	19	C010, C016, C017 and C019.
Time	Event period	2011–2023		Evidence package freeze: 24 June 2026.

Note: Counts describe the frozen purposive corpus. Full source references and evidence locations are provided in the supplementary source matrix.

The unit of analysis was an Oversight-Control Evidence Statement (OCES): one traceable statement that specified, demonstrated, limited, or contested an oversight function, enabling condition, or failure. Forty-one OCES records entered the analysis, comprising twenty-two prescriptive requirements and nineteen incident or contrast records. Each record retained a source identifier, evidence location, statement type, O2C coding, uncertainty note, and separate field for formal EU high-risk status. Sources were classified through an evidence hierarchy. Tier A covered official courts, regulators, audits, accident investigations, and commissions. Tier B covered peer-reviewed empirical studies or method-explicit institutional research. Secondary scientific analysis supported interpretation but did not substitute for the primary case record, and discovery-only material was excluded from strong claims. Inclusion required a materially consequential automated role, an identifiable human oversight role, and evidence relevant to observation, interpretation, judgment, intervention, escalation, or recovery. The complete source matrix, OCES dataset, and exclusion log are supplied in the supplementary package. The OCES structure separated epistemic status from analytical interpretation. Normative requirements, declared capabilities, demonstrated capabilities,

documented failures, contrast cases, and hypotheses could therefore not be silently merged. It also prevented the whole document from becoming the unit of coding: a single framework could yield several independently traceable requirements, while a case could support only the functions explicitly evidenced in its record.

3.3 Coding and analytical constructs

Coding followed the six O2C functions and the cross-cutting conditions defined in Table 2. For cases, each function was coded Present, Limited, Constrained, Delayed, Failed, or Not assessed. Prescriptive documents were coded Explicit, Partial, Implicit, or Not detailed. Fifteen failure mechanisms (F01–F15) captured specific control breakdowns, including weak anomaly detection, poor interpretability, automation bias, authority-boundary failure, non-actionable correction, time or workload constraints, missing escalation or fallback, weak logging or reporting, misaligned objectives, inadequate validation, proprietary opacity, workflow mismatch, and organizational governance failure. A targeted rule-stability check re-examined ambiguous or central codes. It produced one case-level revision: C008 was recoded from F04 to F08 because the record documented automated rejection without a human acceptance or rubber-stamping process; the relevant mechanism was missing meaningful review or recourse. F05 was clarified as an authority-boundary failure, and F15 as an organizational governance, incentive, or culture failure. Phase 2 values were preserved in the audit trail, and no inter-coder reliability or delayed self-recoding procedure was claimed.

Table 2
Operational definitions and verified coding rules for the Oversight-to-Control framework.

Seq.	Construct	Operational definition	Allowed case statuses	Primary capability score	Primary interruption rule	Robustness note
1	Observe	Access relevant system state, context, data, outputs and known limits.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.
2	Detect	Recognize anomalies, unsafe states, conflicts, bias signals or limit violations.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.
3	Interpret	Understand the meaning, reliability and likely consequences of an output.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.
4	Challenge & Decide	Form independent judgment, contest, reject or escalate an output.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.
5	Intervene	Alter, pause, reverse, override or stop a decision or action.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.
6	Recover & Learn	Contain harm, restore an acceptable state, investigate and improve the system.	Present; Limited; Constrained; Delayed; Failed; Not assessed	Present = 1; Limited = 0.5; Constrained = 0.25; Delayed = 0.25; Failed = 0; Not assessed=missing	Earliest function at the maximum observed severity within the case.	Limited and Constrained are regrouped in sensitivity F; rankings are compared.

Note: Scores are transparent descriptive aids rather than validated measurement scales.

3.4 Analytical strategy

RQ1 was analyzed by comparing the presence and explicitness of the six O2C functions and four summarized transversal dimensions—competence, authority, workload/time, and logging/feedback—across the eight prescriptive documents. Explicit, partial, implicit, and absent coverage were assigned transparent descriptive weights of 1, 0.5, 0.25, and 0. RQ2 used categorical function profiles and counts of the F01–F15 mechanisms. Cases were grouped into multi-function breakdown, upstream information or interpretation breakdown, authority/escalation and delayed intervention, and partial-control or model/workflow limitation profiles. Mechanisms were also grouped across technical/model, cognitive, information/workflow, and organizational/institutional layers. RQ3 compared normalized prescriptive coverage with case capability scores using Present = 1, Limited = 0.5, Constrained = 0.25, Delayed = 0.25, and Failed = 0; Not assessed values were omitted. These scores were descriptive analytical aids rather than validated scales. All scored results were reported with the underlying categorical counts. Requirement–capability gaps were interpreted only within the bounded corpus. For each case, the primary interruption was defined as the earliest O2C function at the maximum observed severity. A first-degradation field identified the earliest non-Present status, and profile groups were assigned through documented rules rather than post hoc narrative grouping. Mechanism

counts allowed multiple codes per case because the evidence frequently documented interacting technical, workflow, cognitive, and institutional conditions. The analysis did not weight cases by publicity, severity, or number of source pages.

3.5 Robustness and reproducibility

Reproducibility and internal consistency were assessed before interpreting the results. Baseline assertions confirmed the v1.0 versions of the source matrix, case matrix, OCES dataset, codebook, and citation matrix. Identifier uniqueness, source–case–OCES alignment, allowed values, evidence locations, and mandatory exclusions were checked programmatically. The targeted rule-stability review focused on F03, F04, F05, F06, F08, F09, F11, F12, F14, and F15, all mixed cases, records with multiple mechanisms, and observations supporting central claims. Every change was entered in a versioned revision log. This procedure was an internal mono-researcher quality control, not a measure of inter-coder agreement. The analytical dataset was released as v1.1 to make the C008 recoding and derived variables explicit while preserving the frozen source evidence.

Seven sensitivity analyses examined the stability of the findings: (A) the A + B corpus versus Tier A evidence only; (B) inclusion versus exclusion of the three Tier B scientific cases; (C) cases from 2018 onward versus the complete period; (D) leave-one-domain-out analysis; (E) inclusion versus exclusion of the mixed or negative cases; (F) separate versus grouped treatment of Limited and Constrained; and (G) descriptive inclusion of P008, P009, and C003 without admitting them to the main claims. Triangulation compared regulatory frameworks, risk-management guidance, judicial and regulatory decisions, accident investigations, and scientific deployment studies. Negative-case analysis centered on State v. Loomis, the Apple Card investigation, and the diabetic-retinopathy deployment in Thailand. These cases tested whether O2C could distinguish opacity from illegality, partial control from its absence, and workflow fragility from a discrete harmful incident. Robustness was rated Satisfactory rather than Solid because the central patterns remained stable but the design retained public-record and cross-domain limitations. Sensitivity results were evaluated through changes in function-level scores, stability of the leading mechanism set, and case-ranking correlations. The analysis did not treat an unchanged numeric rank as the only indication of robustness; it also examined whether the substantive interpretation of nominal, delayed, partial, or multi-level control changed. The triangulation matrix classified evidence as convergent, complementary, contradictory, or absent rather than forcing all source types into a single evidentiary scale.

3.6 Reproducibility, ethics and data availability

The study analyzed public documents and did not involve human participants, personal data collection, or animal research; ethics approval and consent were therefore not applicable. The supplementary package contains the anonymized source matrix, case matrix, OCES dataset, codebook, rule-stability and sensitivity files, claim–evidence ledger, figures, tables, and executable notebook. Copyrighted standards are not redistributed. The package records generative-AI assistance used for documentary discovery, structural planning, provisional coding support, and artifact production. Source selection, evidence locations, coding decisions, analyses, interpretations, and the final manuscript were subject to human verification, and responsibility remains with the author. An anonymous repository link and persistent identifier remain to be inserted during the final submission audit.

4 Results

4.1 Prescriptive framework coverage

The analytical corpus combined eight prescriptive documents with nineteen incident or deployment cases spanning 2011–2023. Sixteen cases were supported by official primary evidence and three by peer-reviewed empirical studies. The cases covered autonomous mobility, education, employment, financial services, health care, justice or policing or biometrics, and social protection or public administration. The forty-one included OCES records comprised twenty-two normative requirements and nineteen failure or contrast records. Five cases were classified as multi-function breakdowns, seven as upstream information or interpretation breakdowns, three as authority/escalation and delayed-intervention profiles, and four as partial-control or model/workflow limitations. These counts describe the purposive public-record corpus and are not estimates of how frequently the mechanisms occur in a wider population. Table 1 presents the corpus characteristics. Social protection and public administration formed the largest group, with six cases. Autonomous mobility, health care, and justice/policing/biometrics each contributed three; employment contributed two, and education and financial services one each. Three Tier B cases supplied

external-validation or field-deployment evidence rather than official adjudication. Their inclusion broadened the analysis beyond litigated failure while the Tier A-only sensitivity test assessed whether they altered the central pattern.

The prescriptive analysis showed strong convergence around the ability to challenge a system and intervene in its operation or output. Challenge and Decide and Intervene were explicit in all eight documents, each receiving a normalized coverage score of 1. Observe and Detect were explicit in five documents and partial in three, for scores of 0.8125. Interpret was explicit in six and partial in two, yielding 0.875. Recover and Learn was explicit in seven and partial in one, yielding 0.9375. This pattern indicates that the instruments generally recognize oversight as more than passive monitoring. They expect a human or institution to form an independent view and to disregard, reverse, stop, escalate, or otherwise modify an AI-supported process. Table 3 reports the complete coverage matrix, while Fig. 3 displays the explicit, partial, implicit, and not-detailed codes.

Table 3
Coverage of O2C functions and selected transversal conditions across eight prescriptive frameworks.

Document	Obs.	Det.	Int.	Chal./Dec.	Interv.	Rec./Learn	Comp.	Auth.	Work/Time	Log/Feed
EU AI Act	E	E	E	E	E	P	E	E	I	E
NIST AI RMF	E	E	E	E	E	E	E	E	E	E
NIST AI RMF Playbook	E	E	P	E	E	E	E	E	E	E
OECD AI Principles	P	P	E	E	E	E	I	I	–	E
UNESCO Recommendation	P	P	P	E	E	E	E	E	P	E
Council of Europe Convention	P	P	E	E	E	E	I	E	–	E
ICO Human Review	E	E	E	E	E	E	E	E	E	E
EU HLEG Ethics Guidelines	E	E	E	E	E	E	I	I	–	E

Note: E=explicit, P=partial, I=implicit, and –=not detailed. Scores compare specification within the bounded eight-document corpus.

Coverage was less uniform for the conditions that make those functions practicable. Competence was explicit in five documents and implicit in three, for a score of 0.7188. Authority was explicit in six and implicit in two, for 0.8125. Logging and feedback were explicit in all eight. By contrast, workload and time were explicit in three documents, partial in one, implicit in one, and not detailed in three; the resulting score, 0.4688, was the lowest among the summarized transversal dimensions. The documents also differed in legal force and operational granularity. Binding legal provisions and treaty obligations established broad requirements, whereas audit guidance and risk-management frameworks supplied more detailed prompts concerning reviewer resources, caseload, fallback, and monitoring. The result supports CL01: prescriptive convergence is strongest at the level of challenge and intervention, while the temporal and workload conditions of meaningful review are specified less consistently.

4.2 Incident failure mechanisms and cross-case profiles

The case analysis produced four recurring profiles. Five cases displayed multi-function breakdown, defined as failure in at least four of the first five O2C functions. Seven were characterized by an upstream information or interpretation breakdown without meeting that full-chain criterion. Three involved authority, escalation, or delayed intervention even though upstream information was not completely failed. Four retained partial control while revealing model or workflow limitations: State v. Loomis (C010), the Apple Card investigation (C016), the external validation of the Epic Sepsis Model (C017), and the diabetic-retinopathy deployment in Thailand (C019). Table 4 reports each function status, primary interruption, mechanism combination, and evidence level. Figure 4 displays the profiles using the categorical Present, Limited, Constrained, Delayed, and Failed codes,

thereby preserving differences that would be obscured by a binary success/failure label. The five multi-function cases included configurations in which observation, detection, interpretation, judgment, and intervention were all substantially impaired. Upstream breakdown profiles typically retained some recovery capacity but showed that inadequate observation, detection, or interpretation constrained later action. The authority/escalation profile captured cases in which information was at least partly available but institutional discretion, reporting, or recourse delayed control. The four partial-control cases prevented a failure-only interpretation.

Table 4
Cross-case profiles of O2C functions, failure mechanisms and analytical configurations.

Case	Domain group	Tier	Profile	O/D/I/C/I/R	Primary interruption	Failure mechanisms
C001	Education	A	Authority / escalation and delayed intervention	L/D/L/L/D/P	Detect	F03 poor interpretability; F08 weak appeal/fallback; F12 inadequate pre-deployment validation; F15 governance pressure
C012	Justice / policing / biometrics	A	Authority / escalation and delayed intervention	L/L/L/C/D/P	Challenge Decide	F05 excessive discretion; F09 inadequate documentation; F12 equality-risk assessment gap
C015	Autonomous mobility	A	Authority / escalation and delayed intervention	P/P/L/C/D/F	Recover Learn	F09 incomplete reporting; F10 weak organizational learning; F15 compliance-governance failure
C004	Social protection / public administration	A	Multi-function breakdown	F/F/F/F/F/D	Observe	F04 no meaningful review; F08 weak recourse; F09 inadequate notice/logging; F12 faulty automation; F15 institutional failure
C011	Justice / policing / biometrics	A	Multi-function breakdown	L/F/F/F/F/F	Detect	F02 anomaly-detection failure; F04 automation bias; F09 weak logging; F12 inadequate testing; F15 governance failure
C013	Autonomous mobility	A	Multi-function breakdown	F/F/F/F/F/P	Observe	F02 detection failure; F07 attention/workload; F08 inadequate fallback; F12 validation gap; F15 safety-management failure
C008	Employment	A	Multi-function breakdown	F/F/NA/F/F/P	Observe	F11 discriminatory rule; F08 missing human review/recourse; F12 inadequate compliance testing
C014	Autonomous mobility	A	Multi-function breakdown	L/F/F/F/F/P	Detect	F04 automation overreliance; F07 attention constraint; F08 ineffective engagement monitoring
C016	Financial services	A	Partial control / model-workflow limitation	P/P/L/L/D/P	Intervene	F03 limited consumer-facing explanation; F06 initially constrained appeal; mixed/negative case
C017	Healthcare	B	Partial control / model-workflow limitation	P/L/L/P/P/L	Detect	F02 unreliable detection; F07 alert burden; F12 inadequate external validation
C019	Healthcare	B	Partial control / model-workflow limitation	P/L/L/P/L/P	Detect	F07 workflow/time constraints; F08 exception handling; F14 role/workflow mismatch
C010	Justice / policing / biometrics	A	Partial control / model-workflow limitation	P/P/L/P/P/NA	Interpret	F03 limited interpretability; F13 proprietary methodology; mixed control case

Note: O/D/I/C/I/R denotes Observe, Detect, Interpret, Challenge and Decide, Intervene, and Recover and Learn. P=Present, L=Limited, C=Constrained, D=Delayed, F=Failed, NA = Not assessed. Full source details appear in Supplementary Table S1.

Case	Domain group	Tier	Profile	O/D/I/C/I/R	Primary interruption	Failure mechanisms
C002	Social protection / public administration	A	Upstream information / interpretation breakdown	F/D/F/C/D/D	Observe	F04 rubber-stamping; F05 insufficient discretion/authority; F08 weak escalation; F10 delayed learning; F15 institutional incentives
C020	Social protection / public administration	A	Upstream information / interpretation breakdown	F/D/F/C/D/D	Observe	F05 weak authority/escalation; F08 inadequate review path; F09 opacity; F10 delayed learning; F15 institutional failure
C006	Social protection / public administration	A	Upstream information / interpretation breakdown	L/F/F/C/D/P	Detect	F03 opacity; F08 inadequate notice/appeal; F13 inaccessible logic
C007	Employment	A	Upstream information / interpretation breakdown	L/L/F/F/C/P	Interpret	F03 nontransparent score; F06 non-actionable challenge; F13 proprietary model
C005	Social protection / public administration	A	Upstream information / interpretation breakdown	L/F/L/C/D/P	Detect	F03 opacity; F05 constrained authority; F08 weak appeal; F12 insufficient validation
C009	Social protection / public administration	A	Upstream information / interpretation breakdown	L/L/F/C/D/P	Interpret	F03 opacity; F09 weak transparency; F13 inaccessible risk model
C018	Healthcare	B	Upstream information / interpretation breakdown	L/F/L/L/D/P	Detect	F11 misaligned proxy; F12 insufficient fairness validation; F15 organizational-objective failure

Note: O/D/I/C/I/R denotes Observe, Detect, Interpret, Challenge and Decide, Intervene, and Recover and Learn. P=Present, L=Limited, C=Constrained, D=Delayed, F=Failed, NA = Not assessed. Full source details appear in Supplementary Table S1.

Interpretation and timely intervention were the most consistently degraded functions. Interpret was assessable in eighteen cases; none was coded Present, nine were Limited, and nine Failed. Its mean descriptive capability score was 0.25. Intervene was degraded in seventeen of nineteen cases: two were Present, one Limited, one Constrained, ten Delayed, and five Failed, for a score of 0.2763. Challenge and Decide was degraded in sixteen cases, with only C010, C017, and C019 coded Present. Detect was also degraded in sixteen cases, while Observe was degraded in fourteen. Recover and Learn differed: it was Present in twelve of eighteen assessable cases and had a score of 0.7361. The mixed cases demonstrate that degradation was not synonymous with complete control loss. C010 retained judicial judgment and intervention despite limited interpretability; C016 retained monitoring and review but showed limited explanation and delayed appeal; C017 and C019 retained professional challenge under model or workflow constraints. The pattern was not simply one of deteriorating model performance. Observe was Present in five cases, Limited in nine, and Failed in five. Detect was Present in three, Limited in five, Delayed in three, and Failed in eight. These distributions indicate that an institution could possess some information about a system while lacking an effective process for recognizing the relevant problem. Similarly, the three Present Challenge and Decide codes did not guarantee fully effective oversight: C017 and C019 still contained model or workflow limitations, and C010 retained only Limited interpretation. The chain must therefore be read configurationally rather than as six independent pass/fail indicators.

The mechanism analysis identified missing escalation, fallback, or recourse (F08) as the most frequently coded mechanism, appearing in ten cases. Inadequate validation or monitoring (F12) appeared in nine, and organizational governance, incentive, or culture failure (F15) in eight. Poor interpretability (F03) appeared in seven and weak logging, notice, or reporting (F09) in six. Automation bias or rubber-stamping (F04) appeared in four cases, the same count as authority-boundary failure (F05) and time, workload, or attention constraint (F07). The targeted rule-stability revision affected this ordering only at the margin: C008 moved

from F04 to F08 because no human acceptance process was documented. Figure 2 reports the mechanism counts after that correction. The distribution indicates that the selected oversight failures were not dominated by a single cognitive bias. Escalation design, model assurance, and organizational governance were at least as important as reviewer behavior. Lower-frequency mechanisms remained substantively important. Weak anomaly detection (F02) appeared in three cases; authority-boundary failure, automation bias, and time/workload constraints each appeared in four; proprietary or inaccessible model logic appeared in four; and misaligned objectives appeared in two. F14, role or workflow mismatch, appeared only in C019, but the field study provided unusually direct evidence of how infrastructure and clinical workflow can shape oversight. A zero count for F01 did not mean observability was unimportant: observation limitations were captured in function statuses, while no case met the narrower rule for a distinct F01 mechanism after the stability review.

Mechanisms generally combined across levels. Sixteen of nineteen cases involved at least two of the technical/model, cognitive, information/workflow, and organizational/institutional layers, and seven involved at least three. The Uber Tempe case, for example, combined detection problems, attention and workload constraints, inadequate fallback, validation gaps, and safety-management failure. Benefits and public-administration cases combined opacity, authority limits, weak appeal or notice, delayed learning, and institutional incentives. Scientific health-care cases added model validation, proxy choice, alert burden, and workflow integration (Wong et al. 2021; Obermeyer et al. 2019; Beede et al. 2020). The cross-case configurations also revealed asymmetries. Some cases retained action authority but had limited information; others provided some information while constraining challenge or intervention. In twelve cases, Recover and Learn was Present while Intervene was not Present, indicating that institutional correction frequently followed rather than prevented the adverse trajectory. These combinations support CL04: the relevant unit is a multi-level control configuration, not an isolated human error. Specific configurations illustrate why layer counts matter. The Rite Aid matter combined weak detection, deference to automated matches, inadequate testing, weak logging, and corporate governance deficits. The Uber and Tesla investigations joined system limitations to attention, engagement-monitoring, and fallback failures. Public-benefit cases joined opacity or weak notice to constrained authority, delayed intervention, and institutional learning failures. By contrast, the population-health study linked a misaligned cost proxy to fairness validation and organizational objective setting. These patterns distribute the control problem across the AI system lifecycle and the institutional context in which human action is expected.

4.3 Requirement–capability–evidence gaps

The requirement–capability–evidence matrix compared the normalized coverage of each O2C function with the case-level capability score. The comparison is descriptive and does not assume that the eight frameworks and nineteen cases form representative populations. Its purpose is diagnostic: to identify functions that are prominent in prescriptive instruments but weak in the case evidence, while retaining the raw categorical distributions. Table 5 therefore reports the coverage score, capability score, numerical gap, number of assessable cases, and present or degraded counts. The matrix also includes cross-cutting evidence concerning workload, authority, escalation, logging, and organizational learning. The result is not a validated measurement of compliance. It is a structured way to expose where a formal oversight requirement may lack demonstrated operational capacity. The numerical comparison used the same upper bound for prescriptive coverage and case capability, but the two scores have different evidentiary meanings. Coverage indicates how explicitly a function is specified in the eight documents; capability summarizes the categorical status documented in the nineteen cases. The difference between them is therefore not a compliance rate. It is an index of analytical tension between normative prominence and case-level enactment. This distinction is retained in Table 5 through separate raw counts and narrative evidence patterns.

Table 5
Descriptive gaps between prescriptive coverage and case-level O2C capabilities.

Dimension	Prescriptive coverage	Case capability	Gap	Degraded cases	Interpretation
Observe	0.8125	0.5	0.3125	14	Substantial operational gap
Detect	0.8125	0.3289	0.4836	16	Substantial operational gap
Interpret	0.875	0.25	0.625	18	Substantial operational gap
Challenge & Decide	1.0	0.3289	0.6711	16	Substantial operational gap
Intervene	1.0	0.2763	0.7237	17	Largest prescription-to-capability gap
Recover & Learn	0.9375	0.7361	0.2014	6	Smallest gap, but largely ex-post
Workload and time	0.4688	—	—	4	Weak normative detail and recurring operational relevance.
Authority boundaries	0.8125	—	—	4	Often prescribed, but authority can be absent or insufficiently bounded.
Escalation, fallback and recourse	—	—	—	10	The ability to escalate or fall back is a recurrent operational bottleneck.
Logging, notice and reporting	1.0	—	—	6	Universal normative coverage does not ensure enactment.
Organizational governance and learning	—	—	—	8	Governance failures connect local oversight breakdowns to institutional control.

Note: Coverage and capability scores are transparent analytical aids, not validated scales. Raw categorical counts remain the primary evidence; a dash indicates that a cross-cutting row was not assigned a directly comparable capability score.

Intervene showed the largest descriptive gap. Its prescriptive coverage score was 1, compared with a case capability score of 0.2763, yielding 0.7237. Challenge and Decide followed with a gap of 0.6711: the function was explicit in every prescriptive document but Present in only three of nineteen cases. Interpret had a gap of 0.625, with no Present code among eighteen assessable cases. Detect showed 0.4836, Observe 0.3125, and Recover and Learn 0.2014. The raw counts qualify these values. Ten intervention records were Delayed and five Failed; seven challenge records were Constrained and six Failed; interpretation was evenly divided between Limited and Failed. The gaps therefore reflect different mechanisms. Interpretation concerns the quality and task relevance of understanding, whereas intervention concerns the temporal and practical ability to change the outcome. The scores summarize these patterns but do not replace the categorical evidence. The relative order is more informative than the third decimal place. Intervention and challenge occupy the strongest normative position but some of the weakest case-capability positions. Interpretation is almost as prominent normatively yet was never Present in the assessable cases. Observe has the smallest early-chain gap because several cases retained access to outputs or system state, while Detect falls more sharply because information did not consistently become recognized evidence of a problem. The matrix thus distinguishes mere visibility from diagnostic and action capacity.

Recover and Learn produced the smallest numerical gap, yet its temporal position changes the interpretation. Twelve cases were coded Present, but the recovery often consisted of an investigation, reversal, compensation, revised policy, or learning process after the initial decision or incident. In twelve cases, Present recovery coexisted with an intervention status other than Present. Thus, a comparatively strong recovery score did not establish effective ex-ante or real-time control. It instead showed that institutions sometimes corrected course after operational oversight had failed or acted too slowly. This distinction is central to CL05. A governance arrangement can be accountable after the event and capable of organizational learning while still providing inadequate protection during the decision trajectory. Recovery patterns also varied in quality. A Present code could reflect a court injunction, regulatory settlement, model revision, compensation, an accident investigation, or field-learning

process. These outcomes differ in speed and institutional depth, but all occur downstream of the initial oversight episode. The analysis therefore avoids interpreting the recovery score as evidence that the complete O2C chain worked. Instead, it identifies organizational learning as a separate asset that may reduce future risk while leaving the original intervention gap intact.

Cross-cutting conditions reinforced the same diagnosis. Workload and time had the weakest prescriptive coverage and appeared as F07 in four cases, particularly safety-critical and clinical workflows. Authority-boundary failure appeared in four cases. Missing escalation, fallback, or recourse was the leading mechanism across ten cases. Logging, notice, or reporting failures appeared in six despite universal explicit prescriptive coverage of logging and feedback. Organizational governance failure appeared in eight, while delayed or absent learning appeared in three. These findings show that a formal requirement to review or override an output is insufficient when the reviewer lacks a timely escalation path, when the interface does not make correction actionable, or when the organization does not preserve and use the necessary evidence.

4.4 O2C synthesis, negative cases and robustness

Taken together, the results refine the O2C framework as a diagnostic control chain. Observe and Detect establish whether relevant signals reach the human or institution. Interpret establishes whether those signals can be understood for the task. Challenge and Decide establishes whether an independent judgment is formed. Intervene establishes whether that judgment can change the decision or system state in time. Recover and Learn addresses containment and future improvement but cannot substitute for earlier functions. The prescriptive analysis indicates that governance instruments recognize most of these functions, especially challenge and intervention. The case analysis shows that the corresponding capabilities are frequently limited, delayed, or failed within the selected records. O2C is therefore supported as an analytically developed and corpus-benchmarked framework, not as a validated scale or a claim that every application requires an identical sequence. The framework also makes explicit the difference between functional failure and enabling-condition failure. A function such as Intervene may be Delayed because authority is ambiguous, the interface is non-actionable, the decision window is too short, escalation is unavailable, or the institution discourages deviation. These mechanisms suggest different remedies even when the resulting function status is similar. O2C therefore supports diagnosis at two levels: the location of degradation in the control chain and the mechanism that produced it.

The negative cases tested whether the framework could avoid overclassification. In State v. Loomis, proprietary methodology and limited interpretation coexisted with retained judicial authority and explicit restrictions on the score's use (Wisconsin Supreme Court 2016); O2C therefore records partial rather than absent control. In the Apple Card investigation, limited consumer explanation and an initially inflexible review process did not amount to a regulatory finding of unlawful discrimination (New York State Department of Financial Services 2021); Observe and Detect remained Present, while Interpret and Challenge were Limited and Intervene Delayed. The Thailand deployment documented connectivity, workflow, and exception-handling constraints without a discrete adjudicated harmful incident (Beede et al. 2020); professional judgment remained active, while Detect, Interpret, and Intervene were Limited. These cases support CL07 by distinguishing opacity from illegality, partial control from complete breakdown, and deployment fragility from a catastrophic incident.

The central patterns remained stable under the planned robustness checks. Restricting the corpus to Tier A evidence produced a maximum function-score change of 0.095. Restricting the period to 2018 onward produced 0.083. Leave-one-domain-out analysis produced a maximum change of 0.115, and excluding the mixed cases produced 0.095. Grouping Limited and Constrained yielded a case-ranking Spearman correlation of 0.989 with the primary scoring. Descriptively adding the contextual records changed no function by more than 0.016 and did not alter the leading mechanisms. Some lower-rank ordering changed, particularly among Detect, Challenge, and Intervene, but the broad pattern—strong prescriptive coverage, weak interpretation and timely action, and multi-level mechanism combinations—was retained. Robustness was therefore rated Satisfactory, not Solid, because stability did not remove public-record bias or cross-domain heterogeneity. In the Tier A-only analysis, the leading gap pattern was retained and, for several functions, became more pronounced. Excluding pre-2018 cases produced only a small Interpret/Intervene ordering change. Every leave-one-domain-out run retained missing escalation/fallback among the leading codes, and inadequate validation remained in every leading set. Removing the mixed cases strengthened rather than erased the degradation pattern. The contextual addition of the Canadian directive, ISO public overview, and DUO summary was deliberately excluded from principal claims; its negligible numeric effect confirmed that the scope decision was not driving the results.

5 Discussion

5.1 From nominal oversight to effective control

The findings answer the main research question by showing that human oversight operates as effective safety-control only when a context-appropriate chain of capabilities is available. A human role must receive relevant information, recognize departures from expected behavior, interpret the significance of an output, form an independent judgment, and possess a timely means to change the decision trajectory. These conditions are not reducible to a reviewer's competence or attentiveness. They are jointly produced by the system, interface, workflow, organization, and institutional environment. This conclusion extends critiques of nominal review (Green 2022) and institutional accounts of distrust (Laux 2024) by specifying the functions through which distrust becomes operational. It also complements control-by-design approaches that emphasize feedback and human learning (Te'eni et al. 2026). The selected cases show that formal oversight can coexist with delayed intervention, weak recourse, inadequate validation, or organizational pressure. Effective control is therefore a property of an arrangement, not an attribute of a person placed beside an AI system.

The O2C chain should be understood as a dependency structure rather than a rigid linear sequence. The functions interact, and their relative importance varies by context, but a serious breakdown in one function can constrain the value of the others. Strong intervention authority cannot compensate fully for an anomaly that is unobservable or unintelligible. Conversely, detailed information and explanation do not establish control when the reviewer cannot reject the output or reach an escalation path. Meaningful-human-control scholarship already emphasizes the relation among human reasons, system behavior, and responsibility (Santoni de Sio and van den Hoven 2018; Siebert et al. 2023). The documentary evidence adds a practical implication: those relations must be supported by specific information and action pathways. In this sense, O2C translates a philosophical and design requirement into a diagnostic question for governance—where, exactly, does the capacity to alter the trajectory become unavailable, delayed, or institutionally constrained?

The distinction between intervention and recovery is especially important. Governance frameworks often combine monitoring, incident response, remediation, and organizational learning under a broad accountability or risk-management process. The case evidence shows why these functions should remain analytically separate. Appeals, investigations, compensation, model withdrawal, and policy reform can be essential, yet they occur after a person has lost benefits, received an employment decision, faced a safety event, or experienced another material consequence. Raji et al. (2020) correctly situate audit and corrective action across the lifecycle, while Green (2022) shows that formal human review may not protect affected persons when institutions are poorly designed. O2C adds a temporal test: did the arrangement enable an intervention before or during the harmful trajectory, or only learning after it? A system may be accountable and adaptive over time while failing to provide effective operational control in the case at issue.

5.2 Contribution to meaningful human control

O2C also contributes to meaningful-human-control research by specifying a functional pathway through which tracking and tracing can be enacted. The framework does not replace the normative questions of whose reasons a system should track or who should bear responsibility. Instead, it identifies the operational preconditions for those relationships to affect a decision. Observe, Detect, and Interpret concern the availability and intelligibility of evidence; Challenge and Decide concerns reasoned human judgment; Intervene concerns causal influence over the subsequent action; Recover and Learn concerns the institutional capacity to address failure. The actionable properties proposed by Siebert et al. (2023) and the contextual value trade-offs examined by Boshuijzen-van Burken et al. (2025) are therefore compatible with O2C. The framework adds a common diagnostic vocabulary for comparing different domains without assuming that the same interface, authority model, or timing requirement applies everywhere.

The negative cases further show that control is gradated and configuration-dependent. State v. Loomis demonstrates that limited interpretability can coexist with retained judicial authority and limits on how an algorithmic score is used. The Apple Card investigation shows that transparency and appeal can be deficient even when a regulator does not substantiate the alleged discriminatory outcome. The Thailand deployment shows that technical and workflow friction can limit detection and

intervention while clinicians remain active decision-makers. These distinctions matter because policy debates often oscillate between two over-simple positions: a human is present and therefore control exists, or a system is opaque and therefore control is absent. O2C instead asks which functions are present, which are constrained, and whether the configuration is adequate to the stakes. This approach is consistent with the context dependence identified by Boshuijzen-van Burken et al. (2025) and the need to consider delegation boundaries emphasized by Robbins (2025).

5.3 Human factors and organizational conditions

The mechanism findings qualify human-factors explanations centered on automation bias. Automation bias remains relevant: F04 was documented in four cases and can compromise independent judgment. Yet it was not the leading mechanism. Missing escalation or fallback, inadequate validation or monitoring, and organizational governance failure appeared more often in the selected corpus. This result supports the view that automation bias is contingent on the surrounding decision environment rather than an invariant human tendency (Romeo and Conti 2026; Alon-Barkat and Busuioc 2023). It also cautions against using "human error" as an endpoint. A reviewer's deference may reflect the absence of a meaningful alternative, institutional pressure to follow a score, inadequate training, or a workflow that makes challenge costly. The appropriate response therefore cannot be limited to awareness training. It may require revised authority, safer defaults, staffing, independent escalation, monitoring, or changes to the system objective.

Trust and reliance should be interpreted in the same relational manner. Jacovi et al. (2021) frame trust in relation to vulnerability and an expected contract, while Schemmer et al. (2023) focus on appropriate reliance rather than generalized trust. Zhang et al. (2020) show that confidence and explanation can alter trust calibration without guaranteeing better joint performance, and Bansal et al. (2021) show that explanations can increase acceptance even when they do not improve complementary team outcomes. The O2C findings are consistent with these cautions. A reviewer may be well calibrated about a model's average performance yet unable to recognize the relevant failure in a particular case. Alternatively, the reviewer may identify a problem but lack authority to act. Appropriate reliance is therefore one component of effective control, situated between interpretation and independent decision, but it cannot substitute for organizational actionability.

The interpretation gap likewise suggests that explainability should be assessed by its practical role in control. Miller (2019) and Murdoch et al. (2019) emphasize that explanations are audience- and purpose-relative. Hase and Bansal (2020) and Poursabzi-Sangdeh et al. (2018) show that explanations or transparency do not necessarily improve users' ability to predict errors. Vasconcelos et al. (2023) and Buçinca et al. (2021) demonstrate that engagement and cognitive forcing can help in some conditions, but at a cost. These results align with the case profiles: poor interpretability often combined with non-actionable challenge, inadequate validation, or weak fallback. Interface guidance such as that proposed by Amershi et al. (2019) is valuable when it supports correction, feedback, and recovery. However, an explanation becomes a control resource only when it is timely, task-relevant, and connected to a decision right or intervention mechanism.

5.4 Institutional and societal implications

The prescriptive convergence has institutional implications. The eight frameworks consistently recognize challenge, intervention, logging, and feedback, suggesting a shared normative expectation that human oversight should be active and contestable. Yet the case evidence documents repeated deficits in recourse, validation, authority, notice, and reporting. This divergence is the operational version of the principles-to-practice problem identified in the ethics-governance literature (Jobin et al. 2019). Laux (2024) argues that institutionalized distrust requires democratic design, while Raji et al. (2020) emphasize end-to-end accountability infrastructure. The O2C gap matrix offers a way to make those requirements auditable. Rather than recording that "human oversight" exists, an organization can identify the actor, evidence, timing, authority, intervention mechanism, safe state, escalation route, and learning process associated with each function.

For providers and deployers, this analysis suggests a concrete assurance practice. Before deployment, each O2C function should be translated into a claim and supported by evidence appropriate to the use context. Observation claims may require access to relevant inputs, confidence information, provenance, and operating limits. Detection claims may require monitoring thresholds, exception rules, or independent checks. Interpretation claims should identify the audience and decision task. Challenge and intervention claims should document authority, interface affordances, time budgets, escalation, fallback, and

safe-state behavior. Recovery claims should specify containment, investigation, user recourse, and learning. This approach is consistent with the tools-oriented agenda of Morley et al. (2020), lifecycle governance described by Li et al. (2021), and lifecycle harm analysis proposed by Suresh and Guttag (2021). It also prevents a general statement of human control from standing in for evidence that the pathway works.

The societal relevance extends beyond internal assurance. Effective oversight affects whether an institution can explain and contest decisions, whether affected persons have a meaningful route to correction, and whether responsibility is matched by power. These conditions influence institutional legitimacy and the distribution of technological risk. Steen et al. (2025) call for sociotechnical representations that enable better public conversations about AI, and Shneiderman (2020) argues that human-centered AI should pursue both automation and human control. O2C gives those ambitions an operational form. It makes visible when an organization transfers responsibility to a reviewer without supplying resources, or offers an appeal only after material harm. It also clarifies why principles cannot guarantee ethical practice without institutional implementation (Mittelstadt 2019; Hagendorff 2020). Human oversight is therefore not a narrow interface feature. It is a design problem concerning organizations, rights, and public authority.

5.5 Robustness and transferability

The sensitivity analyses support the internal stability of these interpretations. The principal ordering of functions and mechanisms was retained under stricter evidence tiers, a later time period, exclusion of mixed cases, leave-one-domain-out tests, alternative status scoring, and descriptive inclusion of contextual records. The high correlation between the primary and grouped scoring schemes indicates that the distinction between Limited and Constrained did not drive the case ranking. Nevertheless, stability within the dataset does not resolve selection bias or establish external validity. As surveys of human–AI decision research show, tasks, interventions, and outcome measures vary widely (Lai et al. 2021). The findings should therefore be treated as a robust documentary pattern that warrants prospective testing, not as a universal estimate of oversight performance. The rating of Satisfactory preserves this distinction between internal stability and external generalizability. It authorizes a structured interpretation of the documentary evidence and progression to manuscript drafting, but it does not imply that the framework's sequence, scores, or mechanism frequencies have been validated in operational populations.

6 Limitations

The first limitation is the public-record and purposive nature of the corpus. Cases enter official, judicial, regulatory, investigative, or scientific records because a dispute, failure, or evaluation has become visible. Effective oversight that prevents harm may leave a thinner documentary trace. Conversely, prominent failures may generate unusually rich evidence. The mechanism counts therefore describe the selected records and cannot be interpreted as prevalence estimates. The design improves traceability and diversity of evidence, but it does not support population-level conclusions about how often a particular O2C function fails. Public records also differ in the detail they provide. Accident investigations can reconstruct timelines and organizational controls, whereas settlements may preserve allegations and remedies without adjudicating every fact. The evidence hierarchy and uncertainty notes limit this problem, but they cannot make all cases equally rich.

Second, the cases span different technologies, jurisdictions, legal postures, decision speeds, and levels of autonomy. A judicial review of a risk score, a benefits appeal, a clinical alert, and a safety-driver takeover do not require identical control architectures. The analysis managed this heterogeneity through a constant functional vocabulary, explicit evidence tiers, domain grouping, and leave-one-domain-out sensitivity analysis. These controls support comparison of mechanisms, not interchangeability of contexts. Domain-specific standards and prospective studies are required to determine appropriate time budgets, competencies, interfaces, and intervention powers. The constant O2C dimensions facilitate comparison, yet they may underrepresent domain-specific functions. For example, safe-state management has a different operational meaning in automated driving, clinical work, and public administration. The framework should therefore be adapted through domain-specific subcriteria rather than used as a universal checklist with identical thresholds.

Third, coding was performed within a single-researcher workflow. The targeted rule-stability check revisited ambiguous definitions, mixed cases, and observations supporting central claims, and all changes were versioned. It should not be

interpreted as inter-coder reliability. The normalized coverage and capability scores are also descriptive analytical aids rather than validated measures. Their purpose is to make coding assumptions transparent and compare patterns; the raw categorical distributions remain the primary evidence. Sensitivity analyses reduce dependence on particular scoring choices but do not validate the scale. A second researcher could reasonably draw a different boundary between closely related mechanisms such as interpretability and proprietary opacity, or authority and interface actionability. The codebook, negative cases, and revision log make such disagreements inspectable, but they do not eliminate interpretive judgment.

Finally, the study does not assign formal EU high-risk status to cases without the necessary legal facts and applicable dates. High impact and legal high risk remain separate fields. O2C was analytically developed and evaluated against a structured documentary corpus; it was not externally validated through observation, experimentation, or organizational implementation. Future research should test its usability, completeness, and predictive or diagnostic value prospectively with reviewers, affected persons, providers, and oversight institutions. The study also does not compare alternative oversight designs prospectively, measure affected persons' experiences, or determine whether the presence of a complete O2C configuration reduces harm. Those questions require field, experimental, and longitudinal designs.

7 Conclusion

Human oversight becomes an effective safety-control function when a context-appropriate chain of observation, detection, interpretation, independent judgment, intervention, recovery, and learning is supported by adequate information, competence, time, authority, and actionability. The documentary analysis showed a marked difference between the prominence of oversight in governance frameworks and the capabilities documented in public cases. Challenge and intervention were explicit across all eight prescriptive documents, yet interpretation was degraded in every assessable case and timely intervention in seventeen of nineteen. Formal assignment of a human role is therefore an insufficient indicator of control. The central governance task is not to append a human approval step, but to construct and maintain the conditions under which human judgment can matter.

The article contributes the six-function O2C framework, a mechanism-based failure typology, and a requirement–capability–evidence matrix. Together, these tools shift assessment from the question "Is a human in the loop?" to a more precise inquiry into what the human and institution can observe, understand, decide, change, and learn. They also show why oversight failures are generally multi-level and why recovery after harm should not be confused with timely operational intervention. The typology identifies distinct remedies rather than attributing breakdowns generically to human error: improve observability and validation, redesign interfaces, clarify authority, provide fallback and recourse, preserve records, and align organizational incentives with challenge and learning.

O2C is supported, qualified, and limited by the structured corpus and robustness analyses. It distinguishes total breakdown from partial control and workflow fragility, but it is not a universal or externally validated model. Its practical implication is nonetheless direct: high-impact AI should be deployed only when oversight is designed and evidenced as a socio-technical control capacity. Prospective domain-specific evaluation should now determine how that capacity can be measured and maintained in operation. Such evaluation should examine not only whether reviewers can override an output under test conditions, but whether they can recognize the need to do so, act within the real decision window, and trigger recovery when control is lost. That evidence would allow future work to determine where O2C requires refinement and where it transfers across domains.

Declarations

Funding: No funding was received for conducting this study or preparing the manuscript.

Competing interests: The author declares no relevant financial or non-financial interests.

Ethics approval: Not applicable. The study analyzes publicly available documentary and scientific materials and does not involve human participants or animals.

Consent to participate and publish: Not applicable.

Author contributions: Conceptualization, methodology, formal analysis, investigation, data curation, writing - original draft, writing - review and editing, visualization, and project administration: the author.

Data availability: The author-generated source matrix, incident matrix, OCES dataset, codebook, audit logs, analytical notebook, tables, and figures will be deposited at https://osf.io/dyc87/overview?view_only=2d015a9d82874d198273fddd8b3e9be0. Copyrighted source documents and standards are not redistributed. ISO/IEC 42001 was retained only as a public-overview benchmark and did not support clause-level findings.

Use of generative AI and AI-assisted technologies: Generative AI tools were used to assist documentary discovery, structural planning, provisional coding support, drafting support, and production of analytical and manuscript artifacts. The author verified the source references and evidence locations, reviewed and approved the coding and analytical decisions, checked the reported results against the frozen datasets and notebooks, and assumes full responsibility for the manuscript. No AI system is listed as an author.

Supplementary information: The supplementary package contains the anonymized source matrix, incident and deployment case matrix, OCES dataset, codebook, failure taxonomy, full-text reference verification log, rule-stability and sensitivity files, triangulation matrix, negative-case log, claim–evidence ledger, figures, tables, and executable notebook.

References

1. Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: automation bias and selective adherence to algorithmic advice. J Public Adm Res Theor. https://doi.org/10.1093/jopart/muac007
2. Amershi S et al (2019) Guidelines for human–AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York. https://doi.org/10.1145/3290605.3300233
3. Bansal G et al (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445717
4. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376718
5. Boshuijzen-van Burken C, de Vries M, Munyasya A (2025) Autonomous military systems beyond human control: putting an empirical perspective on value trade-offs for autonomous systems design in the military. AI Soc 40:2507–2523. https://doi.org/10.1007/s00146-024-02000-3
6. Buçinca 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. https://doi.org/10.1145/3449287
7. Council of Europe (2024) Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225. Council of Europe, Strasbourg
8. European Commission High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels
9. European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union
10. 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
11. Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30:99–120. https://doi.org/10.1007/s11023-020-09517-8
12. Hase P, Bansal M (2020) Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2005.01831

13. Information Commissioner's Office (ICO) (n.d.) Data protection audit framework: artificial intelligence – human review. ICO, Wilmslow
14. Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.48550/arXiv.2010.07487
15. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2
16. Lai V, Chen C, Smith-Renner A, Liao QV, Tan C (2021) Towards a science of human–AI decision making: a survey of empirical studies. https://doi.org/10.48550/arXiv.2112.11471. arXiv:2112.11471
17. Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. AI Soc 39:2853–2866. https://doi.org/10.1007/s00146-023-01777-z
18. Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B (2021) Trustworthy AI: from principles to practices. https://doi.org/10.48550/arXiv.2110.01167. arXiv:2110.01167
19. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
20. Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. https://doi.org/10.1038/s42256-019-0114-4
21. Morley J et al (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26:2141–2168. https://doi.org/10.1007/s11948-019-00165-5
22. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116:22071–22080. https://doi.org/10.1073/pnas.1900654116
23. National Institute of Standards and Technology (NIST) (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. NIST, Gaithersburg
24. National Institute of Standards and Technology (NIST) (n.d.) AI RMF Playbook. NIST, Gaithersburg
25. New York State Department of Financial Services (2021) Report on Apple Card investigation. State Department of Financial Services, New York, New York
26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453. https://doi.org/10.1126/science.aax2342
27. OECD (2024) OECD principles on artificial intelligence, updated May 2024. OECD, Paris
28. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan J, Wallach H (2018) Manipulating and measuring model interpretability. https://doi.org/10.48550/arXiv.1802.07810. arXiv:1802.07810
29. Raji ID et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372873
30. Robbins S (2025) What machines shouldn't do. AI Soc 40:4093–4104. https://doi.org/10.1007/s00146-024-02169-7
31. Romeo G, Conti D (2026) Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. AI Soc 41:259–278. https://doi.org/10.1007/s00146-025-02422-7
32. de Santoni F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Rob AI 5:15. https://doi.org/10.3389/frobt.2018.00015
33. Schemmer M, Kühl N, Benz C, Bartos A, Satzger G (2023) Appropriate reliance on AI advice: conceptualization and the effect of explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. https://doi.org/10.1145/3581641.3584066
34. Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe and trustworthy. Int J Human–Computer Interact 36:495–504. https://doi.org/10.1080/10447318.2020.1741118

35. Siebert LC et al (2023) Meaningful human control: actionable properties for AI system design. AI Ethics. https://doi.org/10.1007/s43681-022-00167-3
36. Steen M, Timan T, Vethman S (2025) We need better images of AI and better conversations about AI. AI Soc 40:3615–3626. https://doi.org/10.1007/s00146-024-02101-z
37. Suresh H, Guttag JV (2021) A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305
38. Te'eni D, Yahav I, Schwartz D (2026) What it takes to control AI by design: human learning. AI Soc 41:237–250. https://doi.org/10.1007/s00146-025-02401-y
39. UNESCO (2021) Recommendation on the ethics of artificial intelligence. UNESCO, Paris
40. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein M, Krishna R (2023) Explanations can reduce overreliance on AI systems during decision-making. In: Proceedings of CSCW 2023. https://doi.org/10.48550/arXiv.2212.06823
41. Wisconsin Supreme Court (2016) State v. Loomis, 2016 WI 68
42. Wong A et al (2021) External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 181:1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626
43. Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of FAT* 2020. https://doi.org/10.1145/3351095.3372852

Figures

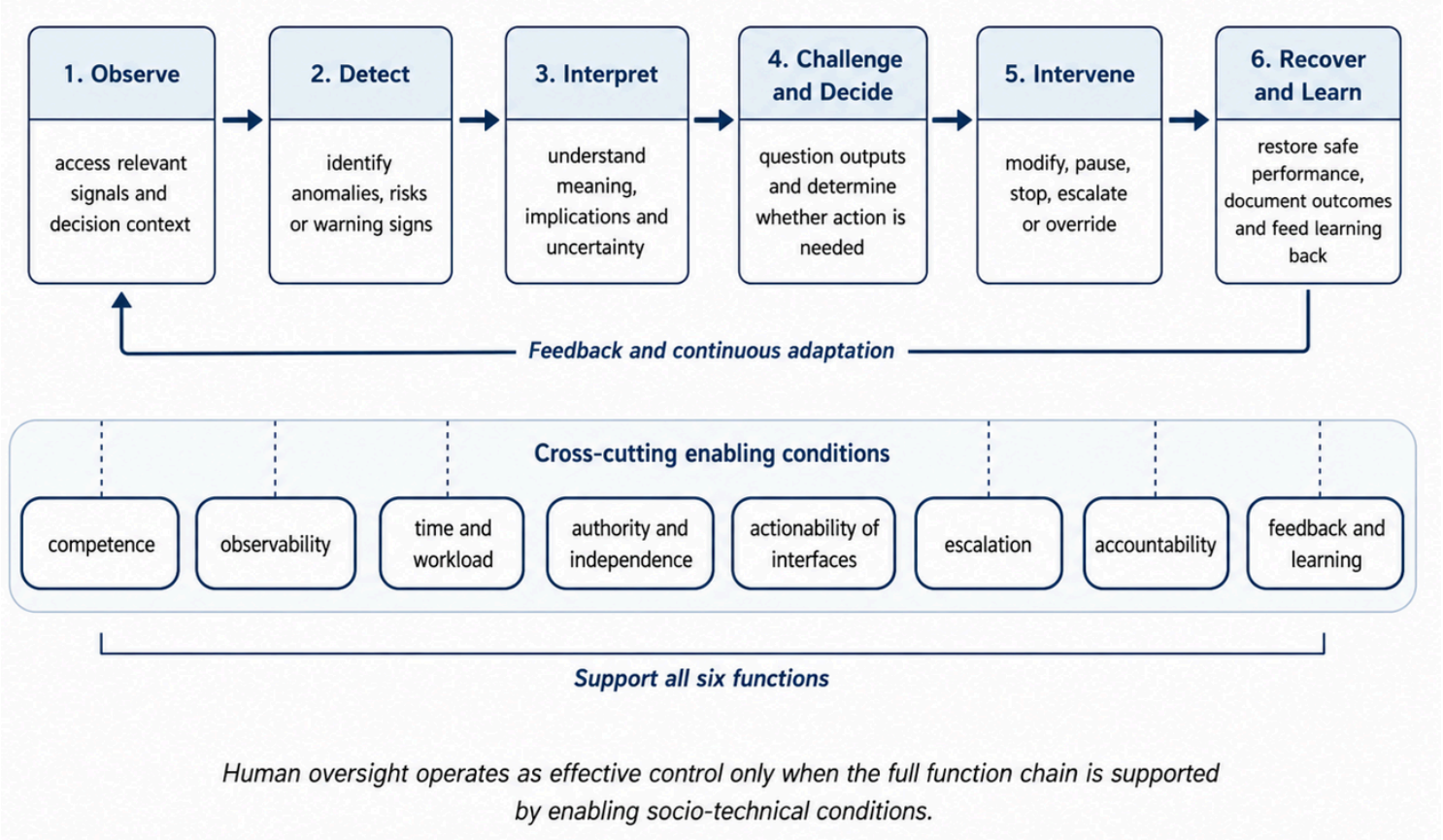


Figure 1

Oversight-to-Control framework showing six functions and cross-cutting enabling conditions.

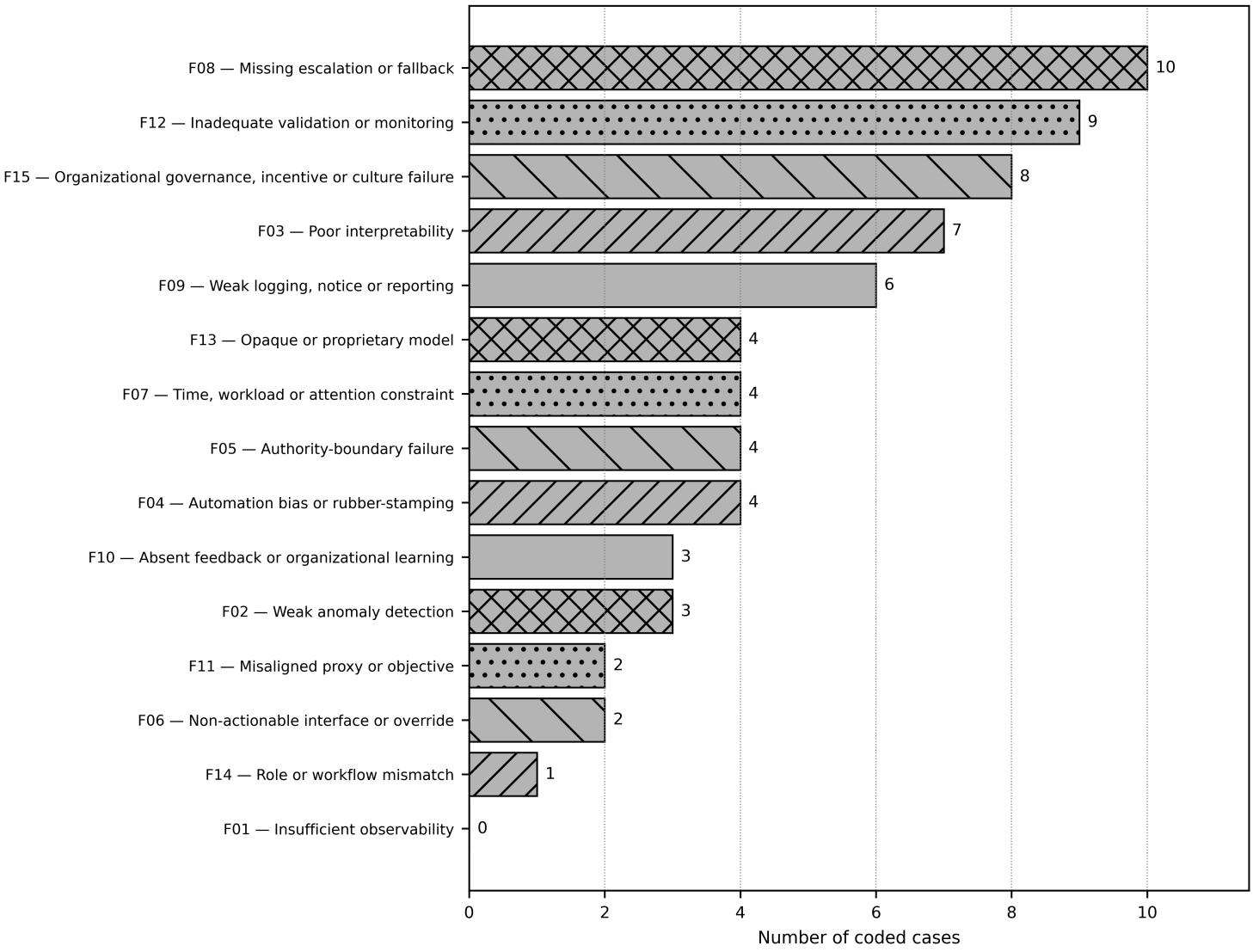


Figure 2

Counts of coded oversight failure mechanisms in the nineteen-case analytical corpus.

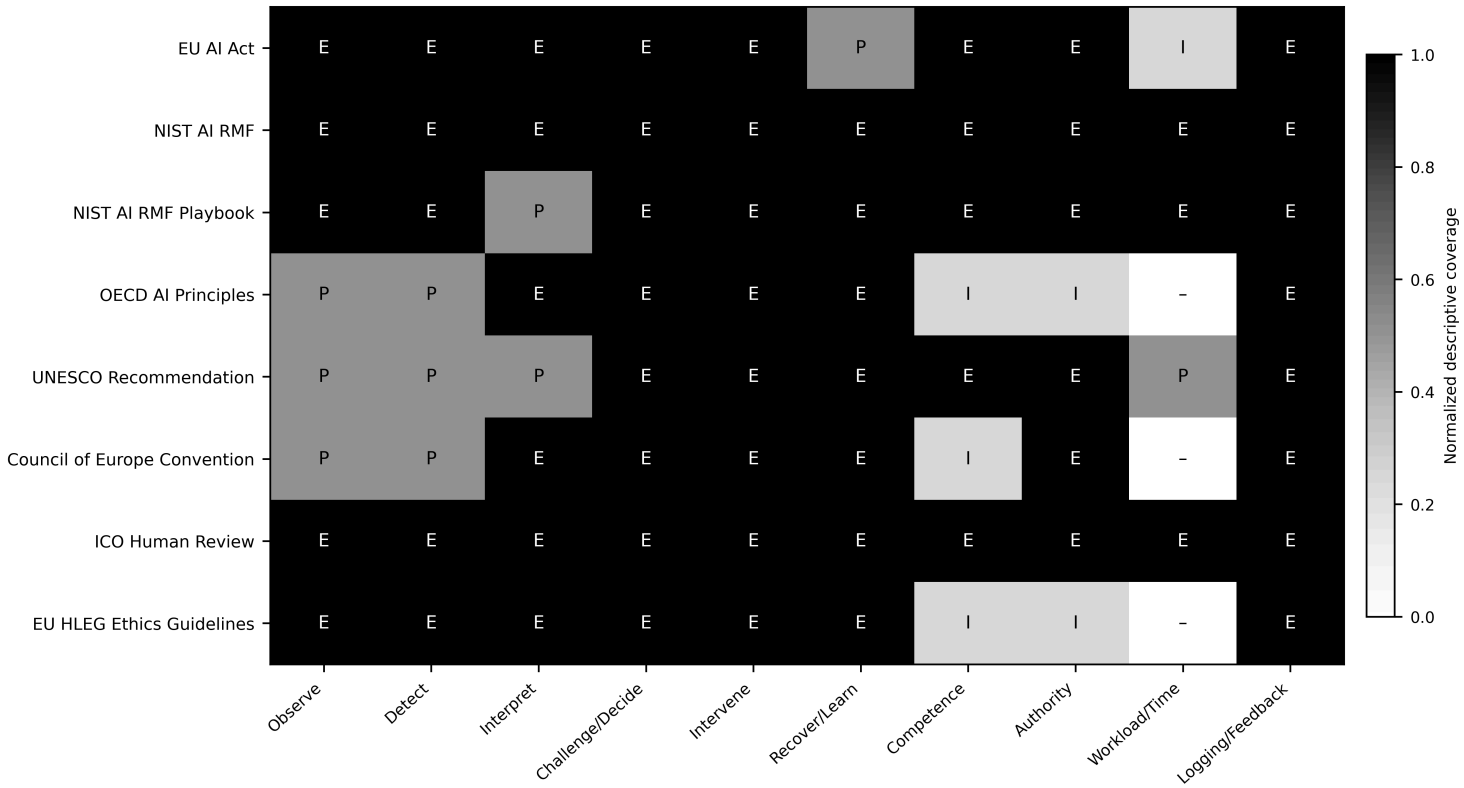


Figure 3

Heatmap of explicit, partial, implicit and absent O2C coverage across prescriptive frameworks.

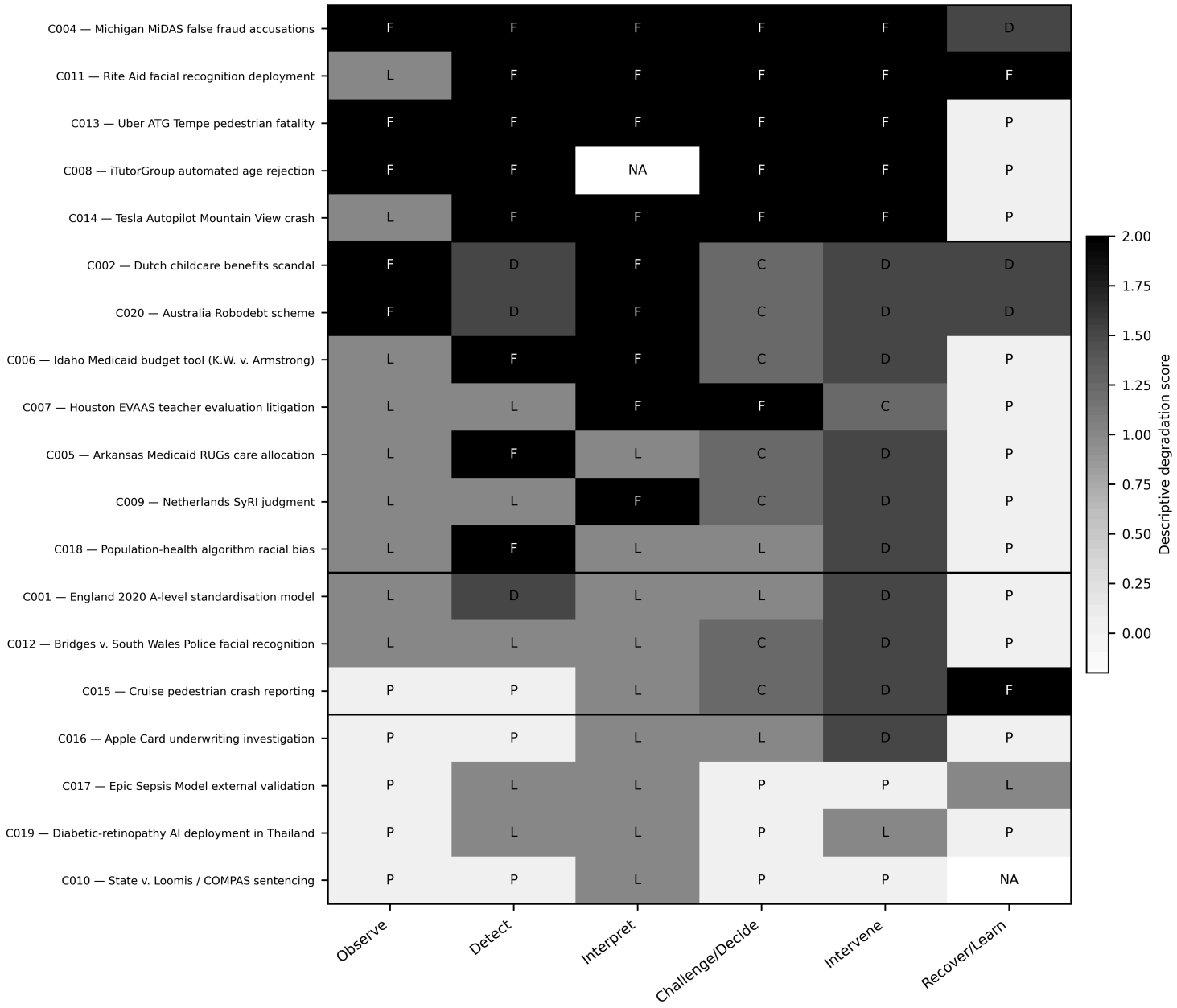


Figure 4

Cross-case profiles of O2C function status, ordered by analytical failure profile.