DESIGNING MEANINGFUL HUMAN OVERSIGHT IN AI

Liming Zhu[1,2], **Qinghua Lu**[1,2], **Ming Ding**[1], **Sunny Lee**[1], **Chen Wang**[1,2]
[1]Data61, CSIRO, Sydney, Australia
[2]School of Computer Science and Engineering, UNSW, Sydney, Australia
firstname.lastname@data61.csiro.au

ABSTRACT

Human oversight is central to safe and responsible AI, but current approaches risk either collapsing agentic AI into mere automation, stripping it of its agentic character, or reducing human agency to a rubber stamp. This paper proposes a design framework that treats agency as layered: AI operative agency in task execution, and human evaluative agency in verification, steering, and substitution. Instead of demanding low-level explanations and controls over how a complex AI model works internally (i.e. internal reasoning faithfulness), we focus on high-level explanations tied to external criteria and human expert understanding (external reasoning faithfulness). This approach retains AI's operative agency while strengthening human's evaluative agency. We also exploit the solve–verify asymmetry by designing AI outputs so that humans can efficiently check and contest them without having to re-solve the task. We present a catalogue of mechanisms (structured rationales, reasoning traces, confidence signals, policy attribution, circuit breakers, appeal bundles) and four end-to-end patterns for practical oversight. We also outline evaluation criteria for AI's agency, human's agency, and joint system agency. The framework provides engineers, safety teams, users, and organisational leaders with a concrete way to design meaningful and effective oversight that preserves accountability while allowing AI to retain its agentic features and autonomy.

***K*eywords** Artificial Intelligent · AI · Agentic AI · Human oversight · Human agency

1 Introduction

Powerful AI and agentic systems, especially generative models often treated as black boxes due to limited explainability, are increasingly deployed in critical domains and high-risk use cases where their outputs shape individual lives, organisational outcomes, and public policy. Calls for "meaningful human oversight" are now common across technical, ethical, and legal frameworks, from industry safety standards to the European Union's AI Act. Yet what counts as meaningful and effective oversight remains under-defined and is often reduced to a checklist of the mere presence of different oversight measures [1]. In practice, oversight often fails in one of two ways. Either humans act as rubber stamps, approving AI outputs they do not fully understand, or systems are constrained so tightly that AI collapses into rule-following automation, losing the very agentic qualities that make AI highly useful [2].

This paper begins from the observation that human agency and AI agency are not a zero-sum game. By agency we mean a system's genuine capacity to steer outcomes toward a goal [3]. Adding human involvement does not necessarily strip agency from the AI. The key is to distinguish **operative agency**, the AI's capacity to generate solutions, from **evaluative agency**, the human's capacity to understand, judge, verify the AI's solution, and, where necessary, steer its outcomes or substitute them with alternative solutions. We use the term evaluative rather than oversight to emphasise that the human role is not merely supervisory or symbolic but actively involves assessment and contextual judgement. For example, an AI system exercising operative agency may generate a highly accurate diagnosis or treatment plan by processing imaging, lab results, and the patient's full medical history. The human doctor's role, grounded in evaluative agency, is not to replicate this diagnostic process nor to surpass the AI's technical accuracy. Instead, it is to understand and evaluate whether the AI's rationale is consistent with what a responsible human expert would find explainable and justifiable, and to judge its appropriateness in light of nuanced context and value-laden considerations — such as treatment burden, patient preferences, or ethical standards. In this way, evaluative agency provides accountability and higher-order judgement that anchor the AI's operative agency within acceptable professional and societal boundaries.

Excessive control and intervention powers without understanding and evaluative agency are of little value: if humans lack the ability to make sense of when and why to intervene, their role reduces either to symbolic oversight, signing off without adding substance, or to blunt overwriting of AI outputs, which risks discarding correct decisions and degrading overall performance. Designed well, these two forms of agency can coexist: AI systems retain their operative agency in "doing" the work, while humans exercise evaluative agency through improved understanding that preserves both accountability and effectiveness.

The main contributions of this paper are: (1) conceptual frame for oversight grounded in agency: how different kinds of understanding reallocate perceived agency between AI and humans; (2) a catalogue of oversight mechanisms and four end-to-end example oversight patterns that engineers and policymakers can adapt across contexts.

The remainder of the paper is organised as follows. Section 2 introduces the conceptual foundations. Section 3 discusses goals and principles for designing oversight. Section 4 presents a catalogue of oversight mechanisms and Section 5 introduces four design patterns. Section 6 discusses related work. Section 7 concludes the paper.

2 Conceptual Foundations: Agency, Understanding, and Solve-Verify Asymmetry

2.1 Agency

Agency is a system's genuine capacity to steer outcomes toward a goal. Following Barandiaran et al. [4], agency can be described through four conditions: individuality (the presence of a system boundary), source of action (causal origin of outputs), goal-directedness (pursuit of norms or objectives), and adaptivity (the ability to adjust behaviour in light of feedback). Abel et al. [3] extend this account by showing that each condition is frame-dependent: whether an AI system is judged to have agency depends on the chosen reference frame—how boundaries, goals, causes, or adaptivity are defined. Unlike humans, whose agency is typically taken as given, an AI may or may not be considered an agent depending on the observer's framing, particularly in relation to the human's role within a socio-technical system.

Building on this frame, we distinguish between two complementary forms of agency: operative agency and evaluative agency. Operative agency refers to the AI's capacity to "do the work": generating solutions in a relatively autonomous fashion. Evaluative agency refers to the human role to understand, judge, verify the AI's solution, and, where necessary, steer its outcomes or substitute them with alternative solutions. Far from being mutually exclusive, these forms of agency work in tandem: operative agency allows the AI to generate solutions without unnecessary human encumbrance, while evaluative agency provides the accountability, contestability, and normative alignment that anchor those solutions in practice.

2.2 The types of understanding

A central issue in designing oversight is what counts as "understanding." Full mechanistic insight into the internal computations of modern generative or learning models is not yet feasible [5]. Building on the established notion of reasoning faithfulness [6], we distinguish between two forms. Internal reasoning faithfulness refers to human-understandable explanations that mirror the model's low-level computations, such as algorithmic circuitry, activation patterns, or concept representations across parameters [7]. External reasoning faithfulness, sometimes described as plausibility, refers to explanations (whether articulated by the AI or imposed afterward) that align with external criteria, established standards, and expert human judgement [8]. Due to limitations in current mechanistic interpretability, understanding via internal reasoning faithfulness is often not feasible. Even when achievable, it may reveal aspects of an AI system's inner workings that are too detailed or complex to be operationally useful for oversight. Understanding and effective oversight therefore tends to rely more on external reasoning faithfulness, which provides interpretable signals that humans can assess, contest, and hold accountable.

2.3 Solve–verify asymmetry

Oversight benefits from recognising the solve–verify asymmetry: solving and verifying are distinct tasks, and their relative difficulty varies by context. In many settings, solving is demanding while verification is comparatively straightforward; in others, generating a solution is trivial but verification is resource-intensive. This asymmetry matters both at design time (when systems are designed, implemented and prepared for deployment) and at runtime (when particular instances are solved and evaluated) (Table 1). Effective oversight design must exploit situations where verification is easier, while mitigating cases where verification is the bottleneck.

Design-time asymmetry

At design time, the "solve" task involves creating the system architecture, specifying objectives, implementing the system, and designing verification tasks and thresholds that will apply across many future instances. The "verify" task is the evaluation of these artefacts before deployment, through requirements review, architecture and design evaluation, testing, and other forms of verification and validation.

There are cases where solving is substantially harder than verifying. For example, developing a high-assurance information-extraction service requires careful schema design, integration of retrieval components, and calibration of confidence signals. Once built, however, verification can be comparatively tractable: artefacts can be stress-tested through coverage checks, constraint validation, and citation-grounding audits that demonstrate whether the system will support efficient instance-level verification later.

Conversely, there are design contexts where solving is easy but verification is difficult. Building an AI chatbot endpoint for open-ended text tasks requires relatively little design effort. Yet verifying that such a system will be safe, factual, fair, and legally compliant across domains is far more challenging, since specifications are open-ended and changing. Likewise, creating a baseline recommender system is straightforward, but verifying its long-term effects on fairness, subgroup outcomes, and feedback loops is demanding.

Runtime asymmetry

At runtime, the asymmetry plays out at the level of individual instances. Here, the "solve" task is the system exercising operative agency to generate a particular solution, while the "verify" task is a human or another system exercising evaluative agency to decide whether to accept, adjust, or contest the solution.

Many instances illustrate solving as harder than verifying. Producing a mathematical proof or factoring a large number may be computationally difficult, yet verification can be decisive by checking each step or multiplying the proposed factors. Similarly, an AI may generate structured outputs such as JSON schemas or citation lists, which can be validated automatically against predefined rules. In these situations, even hallucination could be understood as exploratory generation, acceptable so long as oversight can efficiently separate valid from invalid outputs.

Other instances reverse the asymmetry: solving is easy, but verifying is hard. A generative system can produce a long, persuasive essay or policy draft in seconds. Yet verifying its factual fidelity, completeness, and normative alignment may require hours of human expert review. In such cases, hallucinations are dangerous not because generation is inexplicable, but because verification is costly and prone to oversight gaps.

Implications for oversight

The insight is that verification should be engineered into the system itself but can be done at design time and/or runtime. When verification is generally easier and more reliable at the system level than at runtime, so as much verification as possible should be performed during design. Effective design should therefore reduce or even eliminate the need for costly runtime verification, relying instead on lightweight monitoring, sampled verification and audit. When design-time verification is itself difficult, runtime must be supported with stronger signals, such as confidence indicators, rationale traces, and citations, so that evaluative agency has reliable cues for oversight.

Table 1: Solve–Verify Asymmetry

Stage	Solve harder / Verify easier	Solve easier / Verify harder
Runtime	Hard problems (e.g., proofs, factoring, structured outputs); verification via executing proof steps, tests, rule checks.	Easy generation (e.g., essays, policy drafts, translations); verification for accuracy, fidelity, nuance is costly for high-risk use cases.
Design-time	Complex system design (e.g., high-assurance extraction, code assistant with tests); verification via high quality test suites.	Simple system setup (e.g., chatbot endpoint, baseline recommender); verification for safety, fairness, compliance is complex.

2.4 Oversight of subjective tasks

In subjective domains, disagreement should be treated as a feature rather than a flaw of oversight. Rather than collapsing diverse perspectives into a single answer, effective oversight may require surfacing and preserving multiple viewpoints or counterfactuals for human evaluators to weigh in context. This approach acknowledges the inherent plurality of values and interpretations and strengthens accountability by making trade-offs explicit. At the same time, it is important to recognise the boundary of delegation, i.e., certain forms of judgment, such as legal sentencing, medical diagnosis, or hiring decisions, carry normative and ethical weight that should not be delegated entirely to AI systems. In such cases, oversight goes beyond verifying outputs and extends to safeguarding human responsibility, ensuring that AI remains a tool for support rather than substitution.

3 Oversight as Agency Allocation: Goals and Principles

3.1 What good oversight must achieve

Oversight should increase human understanding, control and accountability without collapsing AI into rule-following automation. Building on the solve–verify asymmetry, the design goal is to make verification efficient, targeted, and reliable so humans can judge plausibility, detect errors, and intervene when needed. Three outcomes matter:

- **Control.** Humans can pause, redirect, or substitute decisions, but these interventions are timely and well-targeted, not symbolic or unnecessarily intrusive.
- **Contestability.** Decisions reduce to comprehensible reasons with evidence so they can be reviewed and appealed.
- **Competence.** Oversight improves overall performance and fairness rather than degrading correct AI outputs through blunt overrides.

3.2 Using the four agency conditions to shape oversight

The four conditions outlined earlier, individuality, source of action, goal-directedness, and adaptivity, do not only describe agency but also guide how oversight should be designed.

- **Individuality** requires clear boundaries between the operative role of the AI and the evaluative role of the human. The AI can be free to generate solutions without unnecessary interference, while humans remain positioned to evaluate the results at meaningful checkpoints. When these boundaries are blurred, attribution and responsibility become diffuse, leading to a situation where agency seems to lie everywhere and nowhere. Oversight therefore needs explicit handover points, supported by identity scoping and logging, so that AI retains operative autonomy in producing solutions and humans retain evaluative agency in determining when and how to intervene. *At design time, define the system boundary and human handovers; at runtime, record when control actually crossed that boundary.*
- **Source of action** highlights the need to trace causal provenance in ways that distinguish between generative processes and evaluative decisions. AI systems produce solutions through exploratory and often opaque generative pathways. Humans may feel they exercise control by shaping prompts or iterating instructions, but this is an illusion of control over the generative process itself. This is one reason why U.S. copyright law [9] does not recognise AI-generated content as human-authored, no matter how complicated or labour-intensive the prompting may be. Where human oversight matters are not in co-owning the generative process, but in evaluating outcomes: assessing provenance records of prompts, inputs, AI calls, and interventions, and deciding whether the results should be accepted, modified, or rejected. Clear provenance metadata enables evaluative agency without requiring humans to replicate or second-guess the generative pathway. *At design time, ensure provenance can be captured end-to-end; at runtime, inspect the specific causal chain for the case at hand.*
- **Goal-directedness** should be understood as a layered structure in which the AI pursues sub-goals while humans retain authority over top-level objectives. AI systems can explore pathways and produce sub-goals that humans cannot or need not fully oversee in real time. The human role is to exercise evaluative judgment over whether these sub-goals align with higher-level aims and constraints such as legal criteria, fairness targets, or policy goals. Making these objectives explicit and traceable allows humans to judiciously evaluate not just the final outcome but also whether the direction of AI exploration remains aligned with legitimate goals. In this way, evaluative agency complements the AI's operative agency without collapsing it into mechanical rule-following. *At design time, publish objectives, constraints, and trade-off rules; at runtime, test whether the current output and sub-goals align with those declarations.*
- **Adaptivity** further clarifies the distinction between operative and evaluative roles. AI systems may adapt at low levels, such as updating internal weights or refining token-level predictions, in ways that humans cannot meaningfully oversee. Attempting to evaluate every micro-adjustment and learning would be infeasible and counterproductive. Human oversight is better targeted at higher-level adaptations that can be expressed in human-understandable terms: changes to policies, constraints, or significant drifts in system behaviour that occur even without explicit re-training. Tools such as versioning, drift detection, and abstain mechanisms allow humans to evaluate and control major shifts while leaving the AI system free to adjust at the micro-level. This division ensures that AI retains adaptive operative capacity, while humans maintain evaluative control over consequential changes. *At design time, set change controls and drift monitors; at runtime, check whether observed behaviour triggers abstain, rollback, or staged rollout criteria.*

Taken together, these conditions create the scaffolding for non-zero-sum agency. AI systems retain operative agency in performing tasks, while humans, equipped with verification surfaces linked to these four conditions, exercise evaluative agency in deciding when and how to intervene. Table 2 and Table 3 align design-time/system-level and run-time/instance-level evaluations to the four agency conditions.

Table 2: System-level evaluation at design time

Agency condition	AI artefacts to expose	Human evaluative checks	System verification surface and sample metrics
Individuality	Architectural boundaries, component map, handover API contracts	Boundary clarity, role separation, identity scopes, auditability	Boundary registry, handover contract tests, % flows with explicit handovers; include a catalogue of model sub-parts eligible for mechanistic probes and their validation status
Source of action	End-to-end provenance model, event schemas, tool integrations	Completeness and integrity of causal capture, time sync, tamper resistance	Provenance schema conformance, log coverage, replayability; document any internal signals admitted as evidence, with calibration tests and limits
Goal-directedness	Policy graph, objective functions, constraint catalogue, trade-off rules	Fitness of goals vs policy, fairness targets, abstain criteria, parameter ranges	Goal ledger with policy links, counterfactual policy tests, fairness sandbox; mapping from goals to explanation requirements and to any validated internal probes tied to goal-relevant behaviours
Adaptivity	Versioning plan, drift monitors, rollout strategy, rollback/abstain design	Change control thresholds, alert quality, red-team playbooks	Change ledger, drift detection precision/recall, rollout failure rate, time-to-rollback; document which drifts trigger external explanations vs internal probe checks

Table 3: Instance-level evaluation at runtime

Agency condition	AI artefacts to expose	Human evaluative checks	Instance verification surface and sample metrics
Individuality	Instance handover records, actor identities, session context	Were boundaries respected and control transferred at the right points	Instance boundary view, handover timestamps, intervention latency; explanation pack indicates whether any internal probe was used
Source of action	Stepwise provenance of prompts, tools, model calls, human edits	Does the causal chain and any admitted internal signal justify the step	Causal trace viewer, missing-evidence alerts, veto precision, acceptance-with-reasons rate; include lightweight saliency or probe readouts only if validated for this class of instance
Goal-directedness	Goals and constraints applied, rationale linked to policy	Do outputs and sub-goals align with declared objectives and acceptable trade-offs	Goals panel with policy links, disagreement analytics, instance-level fairness delta; counterfactuals and contrastive explanations as part of the explanation pack
Adaptivity	Active versions, drift state, abstain/rollback flags	Is behaviour consistent with current versions; do drift or novelty triggers require action	Drift score, abstain utilisation, rollback correctness; show whether drift evaluation relied on external metrics or internal probes

4 Oversight Mechanism Catalogue

Effective oversight needs concrete mechanisms that expose the right artefacts to humans and place control at the points where it adds value. The catalogue below (Table 4) collects practical controls that teams can implement. Each mechanism states what it is, whether it sits at design time or at runtime, which agency dimension it strengthens, and how it supports the solve–verify asymmetry so AI keeps solving while humans verify and intervene.

Table 4: Oversight mechanisms by layer, agency dimension, and solve–verify role

Mechanism	What it is	Layer	Agency dimension improved	Solve–verify contribution
Boundary registry and handover contracts	Declared system boundary, explicit human–AI handovers, API and role contracts	Design-time, then enforced at run-time	Individuality	Lets AI generate within its zone, lets humans verify boundary crossings and intervene at the right checkpoints
Identity and session scoping	Stable identities for AI/Agents, tools, humans; scoped sessions per case	Design-time and run-time	Individuality	Separates who solved from who verified, enables accountable interventions without touching generation
Provenance capture and causal trace	End-to-end logging of inputs, prompts, tools, model calls, edits	Design-time capability, run-time use	Source of action	Human verifies the decisive step in the chain instead of re-solving the task
Structured rationale schema	Machine and human-readable reasons aligned to criteria, with fields for evidence and uncertainty	Design-time schema, run-time population	Goal-directedness	Human verifies plausibility and policy fit without reproducing the solution
External explanation pack	Compact bundle with rationale, citations, contrastive, confidence, abstain flags	Run-time	Source of action, Goal-directedness	Human verifies outcome and reasons at coarse grain when internals are opaque
Validated mechanistic probe hooks	Limited internal signals with documented validity and limits	Design-time adoption, optional run-time use	Source of action, Adaptivity	Enables targeted steerability where evidence supports it, complements external verification
Confidence and uncertainty calibration	Calibrated scores with thresholds tied to impact	Design-time calibration, run-time routing	Source of action	Routes easy cases to light checks and hard cases to deep verification rather than re-solving all cases
Divergence detection and independent checker	Comparison against a second AI/system, heuristic, or ruleset with alerting	Design-time setup, run-time use	Source of action	Flags cases that need verification when solvers disagree, avoids blind trust
Confidence- and impact-aware routing	Triage that considers uncertainty and consequence	Design-time rules, run-time execution	Goal-directedness	Directs human effort to where verification changes outcomes
Veto, substitution, approve-with-reasons	Concrete evaluator controls with reason capture	Run-time	Individuality, Goal-directedness	Human verifies and then acts; AI remains the solver until a justified override
Circuit breaker and safe stop	Immediate halt on predefined risk triggers	Design-time triggers, run-time action	Adaptivity	Human stops harmful solving without micromanaging generation steps
Goals ledger and policy links	Registry of objectives, constraints, trade-offs, legal hooks	Design-time, referenced at run-time	Goal-directedness	Human verifies alignment of outputs and sub-goals to declared aims
Trade-off disclosure and parameter bounds	Transparent sliders or ranges for cost, risk, fairness	Design-time design, run-time adjustment	Goal-directedness	Human verifies and steers high-level goals without re-engineering the solver

Continued on next page

Table 4 – continued from previous page

Mechanism	What it is	Layer	Agency dimension improved	Solve–verify contribution
Adaptivity ledger: versions and change control	Version history, rollout plan, rollback, change notes	Design-time governance, run-time status	Adaptivity	Human verifies behaviour against the active version rather than internals
Drift and novelty monitors	Statistical and domain monitors with thresholds and alerts	Design-time selection, run-time review	Adaptivity	Human verifies that solving remains within trained regime, triggers deeper checks
Shadow mode and staged rollout	Parallel running and progressive exposure	Design-time plan, run-time execution	Adaptivity, Source of action	Lets AI solve while humans verify deltas before full substitution
Counterfactual and contrastive probes	What-if questions and minimal-change explanations	Run-time	Source of action, Goal-directedness	Human verifies decision boundaries without re-solving the whole case
Appeal bundle packaging	One-click export of reasons, evidence, versions, and provenance	Design-time format, run-time export	Individuality, Source of action	Supports external verification by reviewers and courts
Sampling for post-hoc review	Risk-weighted case sampling to maintain skill and detect drift	Design-time plan, run-time execution	Adaptivity, Goal-directedness	Human verifies a subset to validate that solving remains reliable over time

These mechanisms work as paired moves. At design time (system-level evaluation), the organisation commits to a target level of understanding for each angle, for example medium-grained causality through full provenance capture, or fine-grained change control through validated drift monitors and rollback plans. At runtime (instance-level evaluation), evaluators consume the corresponding instance artefacts: handover records that make boundaries visible, causal traces that identify the decisive step, goal panels that link outputs to policy and fairness criteria, and drift scores that explain behaviour at the time of generation. This pairing increases predictability and steerability without forcing humans to re-solve the task. For example, a structured rationale schema provides coarse-to-medium goal understanding at design time by fixing required fields, then yields instance-specific reasons and citations at runtime for quick checking. Validated mechanistic probe hooks offer fine-grained model-level insight where evidence supports them at design time, then inform a small number of high-stakes run-time decisions without encouraging day-to-day micromanagement. Provenance capture and causal traces commit the system to medium-grained causality at design time, then let evaluators locate and judge the decisive step in a case rather than reconstruct the whole pathway. Read this way, the catalogue is a menu for selecting the angle and level of understanding you want, and for deciding where that understanding is guaranteed in assurance and where it is consumed in operation.

To make effective use of the catalogue, select a minimal set that covers all four agency dimensions. At design time, prioritise boundary and provenance capabilities, the goals ledger, and adaptivity governance. At runtime, prioritise the explanation pack, calibrated confidence with abstain, provenance viewing, and evaluator levers. Where validated mechanistic probes exist, add them as optional fine-grain signals. The result is a workflow in which the system continues to solve, while humans verify efficiently and intervene only when it changes outcomes.

5 End-to-End Oversight Pattern Examples

The following are illustrative pattern examples for document evaluation, drawn from recurring oversight challenges. They capture four common modes of evaluation: review (an artefact judged against subjective criteria), conformance (an artefact tested against objective rules), aggregation (many artefacts synthesised into a coherent whole), and prioritisation (many artefacts analysed and ranked). Each pattern combines a subset of mechanisms into a coherent oversight flow. For each, we identify the problem, list selected mechanisms, and show how design-time system-level evaluation and run-time instance-level evaluation work together so that AI systems continue solving while humans verify and intervene where it matters.

5.1 Pattern 1 – Oversight for criteria-aligned assessment (document A against criteria B)

Context

In many domains, such as research funding, regulatory oversight, or academic publishing, submissions like grant proposals, planning applications, or scholarly papers (A) must be **evaluated against well-defined criteria described in policies, rubrics, or guidelines** (B).

Problem

As evaluation workloads increase and AI systems become more deeply embedded within these processes, several risks arise that may undermine trust, transparency, and accountability. One central concern is the presence of incomplete or missing evidence to substantiate claims of compliance, which can lead to erroneous or unjustified decisions. Another is the tendency for human reviewers to rubber-stamp outputs that appear plausible, without critically examining the underlying reasoning. This dynamic can result in cases where a decision is technically correct but rests on flawed or incomplete justification, referred as the "right answer, wrong reasons" problem. These risks are amplified at scale, where the complexity and volume of evaluations exceed the capacity of humans to manually track, assess, and verify each step of the process.

Solution

Potential mechanisms

- Structured rationale schema aligned to B (one row per criterion: decision, evidence excerpt, policy clause, uncertainty).
- Goals ledger with links to B and versioning of B.
- Provenance capture and causal trace for inputs, retrieval, tools, and human edits.
- External explanation pack (claim–evidence table, citations to B, contrastive, confidence, abstain flags).
- Constrained retrieval that only cites B (or approved sources).
- Divergence detection against a simple rules/rubric checker.
- Counterfactual/contrastive probes for borderline criteria.
- Calibrated confidence with abstain.
- Risk-weighted sampling and post-decision audit.
- Veto, substitution, approve-with-reasons; appeal bundle packaging.

Design-time system-level oversight

Translate B into a goals ledger and a mandatory rationale schema so every decision is traceable to a clause and evidence. Configure provenance to capture the decisive steps and retrieval calls. Calibrate confidence at the criterion-level, and define abstain rules for low evidence or out-of-scope items. Stand up a lightweight rules checker to power divergence alerts. Set audit sampling rates that rise with risk or low confidence. Run shadow evaluations before tightening the rubric.

Runtime instance-level oversight

The system produces a recommendation plus an explanation pack that includes the filled rationale table with citations and a concise causal trace. The reviewer reads the rationale instead of re-doing the assessment, uses contrastive on close calls, and inspects divergence alerts where the rules checker disagrees. A proportion of cases are spot-checked at depth to ensure evidence truly supports each criterion and that reasoning traces are consistent with the cited passages. Reviewer actions (approve, veto, substitute) record reasons and feed the audit queue and appeal bundle.

Solve–verify fit

The AI solves by mapping A to B and assembling evidence-backed reasons; the human verifies claim–evidence links, policy fit, and flagged disagreements, intervening only where it changes outcomes.

Signals to track

Verify time per case; evidence-coverage per criterion; disagreement rate with rules baseline; veto precision; audit reversal rate; incidence of "right answer, wrong reasons"; fairness deltas on protected slices.

Benefits

- **Transparency:** Structured rationales, goals ledgers, and provenance capture make every decision traceable to explicit policy clauses and supporting evidence, enabling clear review and contestation.
- **Accountability:** Divergence detection, calibrated confidence with abstain, and audit sampling ensure systematic oversight, with human interventions and outcomes fully logged for appeals and governance.
- **Efficiency:** Explanation packs and constrained retrieval allow humans to focus on verifying claim–evidence links rather than re-performing the entire evaluation process.
- **Scalability:** Risk-weighted sampling and automation of routine checks enable large-scale evaluation while maintaining oversight quality.

Drawbacks

- **Complexity:** Implementing goals ledgers, provenance systems, and divergence detection increases system and governance complexity, requiring specialized expertise.
- **Cost:** Building and maintaining structured oversight processes, including shadow evaluations and continuous policy versioning, demands significant resources.
- **Verification Burden:** For borderline or novel cases, human reviewers may still need to conduct deep spot-checks, reducing efficiency gains at scale.
- **Signal Dependence:** The effectiveness of oversight relies on the quality of generated rationales, provenance data, and alerts; poor or noisy signals can mislead reviewers.

Known uses

- **AAAI AI-Powered Peer Review System:**[1] AAAI has launched a pilot program that incorporates LLMs to enhance the academic paper review process for the AAAI-26 conference. The LLM generates an additional review alongside traditional human expert evaluations and assists the Senior Program Committee by summarizing reviewer discussions, highlighting key areas of consensus and disagreement among reviewers. All LLM-generated content is thoroughly reviewed by human experts, with comprehensive monitoring and evaluation in place to ensure quality and accountability.
- **Revelo eReviewer:**[2] eReviewer is an LLM-powered platform designed for grant proposal peer review to support criteria-aligned evaluations. It helps reviewers by automatically summarising proposals and providing key highlights mapped to predefined rubrics. Human experts oversee the entire process, reviewing all AI-generated outputs and making the final decisions. Detailed audit trails record every evaluation step to ensure transparency and traceability.
- **Vertesia:**[3] Vertesia leverages LLMs to streamline grant proposal evaluations. The system uses retrieval-augmented LLMs to match proposals with explicit program rubrics, generate initial scores, and produce comprehensive proposal summaries. Human reviewers verify, adjust, or override AI recommendations to ensure decisions are based on expert judgment. Oversight mechanisms include interactive chatbot explanations, clear alignment with program criteria, and full logging of evaluations for auditing and governance.

5.2 Pattern 2 – Oversight for conformance checking (document A against manual/guardrail B)

Context

A completed document (A) is typically evaluated against a conformance reference framework (B), such as a style manual, organizational policy or strategy, or a predefined set of guardrails. This evaluation step is an established part of many workflows and is used to ensure that documents meet expected standards. The document could be authored entirely by humans, generated by AI, or created through a combination of both.

Problem

Because B is often lengthy and nuanced, strictly enforcing it during the drafting phase is unreliable. Moreover, the limited controllability of generative systems makes "write-to-spec" approaches difficult to achieve. As a result, a separate checking step is required to assess conformance to B and, where necessary, recommend minimal edits to bring the document into compliance.

[1] https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/

[2] https://revelosoftware.com/ereviewer/

[3] https://vertesiahq.com/solutions/grant-proposal-review-process

Solution

Potential mechanisms

- Goals ledger separating hard constraints (machine-checkable rules) from soft rubrics (judgement-based norms).
- External explanation pack with per-issue rationales, highlight maps, minimal-change diffs, and links to relevant clauses in B.
- Provenance capture for sources, tools, and exemplar sets consulted by the checker.
- Counterfactual rewrites that show the smallest change needed to pass a soft rubric.
- Divergence detection when A's tone or structure falls outside the exemplar band.
- Calibrated abstain for ambiguous cases requiring human judgement.
- Divergence detection and independent checker: eviewer calibration using a golden set; inter-rater agreement tracking.
- Sampling and post-edit audit; rollback of rubric changes over time.
- Appeal bundle packaging with diffs, rationales, and clause references.

Design-time system-level oversight

Encode B into a two-layer goals ledger. Make hard constraints machine-checkable with precise tests. Express soft rubrics as evaluable checklists and labelled exemplars; set parameter bounds to balance concision, nuance, and voice preservation. Define the explanation pack format so rationales, highlights, and minimal-change diffs are mandatory. Configure linters and exemplar detectors with thresholds and calibration data; define abstain criteria for low certainty. Train reviewers on a golden set and measure agreement.

Runtime instance-level oversight

The checker returns a conformance report and, where suitable, suggested minimal edits. Hard issues present failing checks with exact locations and clause links. Soft issues present exemplar-linked rationales, highlight maps, and smallest-change diffs so reviewers can accept or reject edits while preserving voice. Provenance identifies any exemplar sets or tools used. Divergence alerts trigger when tone or structure moves outside the exemplar band. Ambiguous cases invoke abstain for human judgement. Reviewers accept targeted diffs, reject over-corrections, record reasons, and submit cases to sampling or post-edit audit as required.

Solve–verify fit

The AI checks and proposes minimal fixes; the human verifies conformance and decides which edits to adopt. Generation is not steered in real time; conformance is achieved through a clear, auditable checking step.

Signals to track

Hard-check pass rate; agreement with golden set on soft norms; edit acceptance rate by issue type; inter-rater agreement; time to acceptable draft; proportion of abstain cases resolved without rework; rollback frequency after rubric changes; user-reported voice preservation.

Benefits

- **Transparency:** Explanation packs with rationales, highlights, and clause links make every check and suggested edit auditable and easy to trace back to policy or rubric sources.
- **Consistency:** Goals ledgers and calibrated linters standardize the application of hard rules and soft rubrics, reducing reviewer subjectivity and drift over time.
- **Efficiency:** Minimal-change diffs and calibrated abstain thresholds focus human effort on ambiguous cases, improving throughput while preserving author voice.
- **Accountability:** Provenance capture, rollback tracking, and audit sampling provide a verifiable record of checks, tools, and rubric updates for governance purposes.

Drawbacks

- **Complexity:** Encoding nuanced rubrics into exemplar-based systems and maintaining calibration data require significant design-time effort and specialized expertise.

- **Cost:** Building and sustaining rule linters, exemplar detectors, and audit pipelines demands ongoing investment in tooling and reviewer training.
- **Signal Dependence:** Oversight effectiveness relies on high-quality exemplars, thresholds, and provenance data; poor or biased inputs can mislead both the system and reviewers.
- **Verification Burden:** In borderline or novel cases, reviewers may still need to invest substantial time resolving abstains and contested edits, limiting scalability gains.

Known uses

- **ValidMind Document Checker:**[4] ValidMind's LLM-powered Document Checker evaluates model documentation for compliance with regulatory standards such as model risk management policies and AI governance frameworks. The system automatically maps sections of the document to regulatory criteria, identifies gaps, and provides rationales with source evidence. Human validators remain responsible for interpreting nuanced requirements, confirming or dismissing flagged issues, and finalizing compliance assessments. Every interaction is logged, creating a traceable record of human decisions and AI contributions.
- **Norm AI:**[5] Norm AI provides AI agents that act as an automated compliance reviewer for documents and proposals. These agents analyse content against complex regulations and flag any sections that may violate specific rules or laws, pointing reviewers to the exact regulation clauses at issue.
- **Acrolinx:**[6] Acrolinx's enterprise content governance tool uses AI to scan and score documents against an organization's style guides and compliance standards. It checks writing for clarity, tone, terminology, and adherence to rules, ensuring each document follows the required style and policy guidelines. The system provides suggestions or automated rewrites for flagged issues and links each suggestion to the relevant guideline, helping authors fix content while preserving the intended voice.

5.3 Pattern 3 – Oversight for evidence-grounded synthesis and thematic analysis (from large corpus to report)

Context

In many domains, there is a need to synthesize large collections of documents, such as consultation submissions, research papers, or policy reports, into a single structured output. This synthesis typically includes identifying common themes, highlighting nuances and contrasts, and presenting representative quotes along with supporting evidence. The resulting report often serves as a key input for decision-making, strategy development, or policy formation, making accuracy and comprehensiveness essential.

Problem

When synthesizing hundreds of source documents, several risks arise. Important material may never surface, leading to gaps in coverage. Hallucinated or unsourced claims can undermine the credibility of the synthesis, especially when outputs are produced by or assisted with generative systems. In addition, there is a danger of over-confident generalizations that obscure or ignore minority views, resulting in a misleading or biased representation of the underlying evidence. These challenges threaten both the quality and trustworthiness of the final report.

Solution

Potential mechanisms

- Provenance capture and causal trace for corpus register, scope lock, inclusion/exclusion criteria, and deduplication.
- Provenance capture for every claim: passage-level anchors and permalinks.
- Structured rationale schema for synthesis: per-theme claim–evidence tables with source counts, stakeholder types, and date ranges.
- Trade-off disclosure and parameter bounds: transparent weighting scheme (incl. stakeholder weights) and publish bounds.
- Divergence detection: independent model or heuristic re-summaries; minority-view mining.
- Counterfactual and contrastive probes: contrastive theme extraction and counter-evidence prompts.

[4] https://validmind.com/blog/document-checker-for-regulatory-compliance/

[5] https://www.norm.ai/

[6] https://www.acrolinx.com/

- External explanation pack: quote bundles, source lists, confidence, abstain/low-coverage flags.
- Calibrated confidence and abstain criteria tied to evidence density and source diversity.
- Post-synthesis audit: risk-weighted spot checks; reversal tracking; appeal bundle with all citations.

Design-time system-level oversight

Define the corpus register with frozen inclusion criteria, metadata normalisation, and dedup rules so the boundary of "what can be said" is auditable. Map the report outline to a synthesis rationale schema: each theme must include a concise claim, supporting passages with anchors, source diversity stats, stakeholder coverage, recency window, and explicit counter-evidence. Configure coverage probes: a checklist of sentinel items (expected stakeholders, landmark papers, seminal arguments) that the system must surface or abstain on. Set weighting rules for stakeholder categories and document quality signals; publish them in a goals ledger. Stand up divergence detection via an independent summariser or rule-based sampler to challenge the main synthesis. Specify the explanation pack format and the audit sampling plan that oversamples low-coverage or high-impact themes.

Runtime instance-level oversight

The system returns a draft report plus an explanation pack. For each theme it shows the claim, representative quotes with page-level anchors, a source list with counts and dates, coverage indicators (e.g., "6 of 8 stakeholder types represented"), and counter-evidence. Abstain is triggered where evidence density or diversity is below thresholds. Reviewers verify by checking anchors, running contrastive prompts on contentious themes, and inspecting divergence alerts where the independent summariser disagrees. Spot checks target sentinel items and minority-view capture. Reviewer actions include accept with reasons, revise scope/weights, request targeted re-search, or mark for audit.

Solve–verify fit

The AI solves by organising themes and assembling verifiable evidence; humans verify coverage and balance, test disagreements, and intervene where it changes the conclusions.

Signals to track

Evidence-to-claim ratio; source diversity per theme; sentinel-hit rate; minority-view capture rate; abstain utilisation; audit reversal rate; quote anchor validity; recency distribution; time-to-verify per theme.

Benefits

- **Transparency:** Provenance capture with passage-level anchors and quote bundles ensures every synthesized claim can be traced back to its original source, improving auditability and trust.
- **Coverage:** Sentinel-based checks, stratified sampling, and diversity metrics reduce the risk of missing important documents, stakeholders, or minority viewpoints.
- **Balance:** Weighting schemes and divergence detection prevent over-confident generalizations by highlighting conflicting perspectives and surfacing counter-evidence.
- **Accountability:** Structured rationale schemas, audit sampling, and appeal bundles create a clear, verifiable trail of how conclusions were reached and validated.
- **Scalability:** Automating theme extraction, evidence assembly, and coverage probes allows large-scale synthesis while maintaining systematic human oversight.

Drawbacks

- **Complexity:** Configuring registers, coverage probes, weighting rules, and divergence detection requires significant design-time expertise and ongoing calibration.
- **Cost:** Building and maintaining provenance infrastructure, explanation packs, and audit pipelines demands sustained resources and skilled reviewers.
- **Verification Burden:** Contentious or ambiguous themes may still require intensive manual checking, especially when abstain flags or divergence alerts are frequent.
- **Signal Dependence:** Oversight effectiveness relies on the accuracy of anchors, thresholds, and sentinel definitions; weak or biased inputs can lead to misleading synthesis.
- **Governance Load:** Managing versioning, inclusion criteria, and rollback processes for evolving corpora and policies introduces operational overhead.

Known Uses

- **Consensus:**[7] Consensus is an LLM-based literature search tool that scans a large corpus of 200+ million papers and produces a concise synthesis of findings, with citations for every claim. For each query, it retrieves top relevant papers and uses GPT-based summarisation to answer the question while tracing each insight back to its source.
- **Elicit:**[8] Elicit assists researchers in synthesising findings from academic papers using AI. It can retrieve relevant studies for a given question and extract key points or quotes from each, effectively building an evidence table. It prioritises grounding outputs in the original documents, enabling users to trace claims back to the literature.
- **Scite Assistant:**[9] Scite is a scholarly assistant that leverages a large database of scientific publications and an LLM to answer questions with evidence. It uses an AI model (GPT-based) to read through papers and produce an answer with source-based grounding. Scite's system distinguishes supportive vs. conflicting evidence by analysing how each cited paper discusses a topic, thus highlighting reliable research. The tool's output includes citations to the papers it read, helping users verify claims and assess evidence quality.

5.4 Pattern 4 – Oversight for signal detection, novelty, and prioritisation across streams (from news/reports to ranked issue list)

Context

In many operational settings, organizations must continuously scan high-volume information streams, such as media reports, public advisories, and internal incident reports, to identify emerging issues that warrant attention. These streams are monitored against explicit criteria to determine which items are relevant and to prioritize them for subsequent action. This process supports timely decision-making and effective allocation of resources in dynamic environments.

Problem

Several challenges arise in managing this type of monitoring and prioritization workflow. Important signals may be missed entirely, resulting in low recall and the potential for delayed or inadequate responses. Conversely, overly sensitive detection can lead to alert fatigue, where large volumes of false positives overwhelm reviewers and erode trust in the system. In addition, rankings may be shallow or opaque, providing little insight into why certain issues are prioritized, which hinders accountability and informed decision-making. Over time, concept drift, shifts in the meaning of signals or criteria, can degrade system performance, requiring ongoing recalibration and oversight.

Solution

Potential mechanisms

- Trade-off disclosure and parameter bounds: criteria-to-score mapping with transparent weights and adjustable bounds.
- Provenance capture and causal trace: per-candidate evidence vouchers (snippets, links, timestamps).
- Drift and novelty monitors: "first-seen" novelty detectors; similarity clustering and collapse.
- Divergence detection via second model or heuristic rules.
- External explanation pack per issue: score breakdown, supporting evidence, known gaps, and confidence.
- Confidence- and impact-aware routing; abstain for low evidence or low novelty.
- Sampling for post-hoc review: risk-weighted post-decision sampling; appeal-bundle packaging for escalations.

Design-time system-level oversight

Register streams and enforce freshness windows and dedup logic so scanning remains predictable. Translate "worthy" into a criteria-to-score function with weights for impact, relevance, novelty, and credibility; publish bounds in the goals ledger. Configure novelty detectors and clustering to collapse duplicates. Define coverage monitors: watchlists and sentinels for topics you must not miss. Calibrate divergence detection and abstain thresholds tied to evidence quality. Specify the per-issue explanation pack (score breakdown, evidence vouchers, gaps, and confidence). Set sampling rates by tier to audit both false positives and false negatives.

[7] `https://consensus.app/`
[8] `https://theippo.co.uk/`
[9] `https://scite.ai/`

Runtime instance-level oversight

The system outputs a ranked list of issues. Each entry carries an explanation pack: score components, supporting snippets with links and times, similarity cluster context, novelty flags, and any watchlist hits or misses. Items with low evidence or novelty trigger abstain for human review. Reviewers verify top-ranked items first, check provenance and cluster context, and use counter-example mining to reduce spurious spikes. Disagreements with the secondary checker are investigated. Reviewers can reweight criteria within bounds, escalate, or dismiss with reasons; dismissed items feed the learning set for future precision.

Solve–verify fit

The AI solves by detecting and ranking candidate signals; humans verify the ranking rationale, adjust within transparent bounds, and focus attention where it changes downstream action.

Signals to track

Precision/recall against adjudicated samples; watchlist-miss rate; duplicate collapse rate; reviewer reweight frequency and range; time-to-decision by tier; escalation and reversal rates; drift in score distributions over time.

Benefits

- **Transparency:** Explanation packs with score breakdowns, provenance vouchers, and evidence links make rankings auditable and justify why each issue was surfaced or prioritized.
- **Precision and Recall:** Counter-example mining, watchlist sentinels, and novelty detection reduce missed signals while minimizing false positives and alert fatigue.
- **Adaptability:** Adjustable scoring weights and divergence detection allow the system to evolve with shifting criteria and changing data sources.
- **Accountability:** Sampling, escalation bundles, and audit trails ensure that both detections and dismissals are reviewable, supporting governance and trust.

Drawbacks

- **Complexity:** Configuring scoring logic, novelty detectors, clustering, and divergence rules requires specialized expertise and careful ongoing calibration.
- **Cost:** Maintaining registries, provenance tracking, and post-decision audit infrastructure demands sustained resources and operational effort.
- **Verification Burden:** Ambiguous or novel cases may still require intensive manual review, especially when abstains or disagreements are frequent.
- **Signal Dependence:** Performance relies heavily on the quality and stability of streams, watchlists, and thresholds; weak or drifting signals can undermine accuracy over time.

Known Uses

- **Primer Command** [10] is an AI-driven situational awareness platform that monitors over 60,000 global news and social sources in real time. It uses LLM to eliminate duplicates, extract entities (people, places, organizations), and provide structured event updates with "human-grade" precision. The system separates signal from noise, flagging emerging developments or misinformation spikes and presents them on a single dashboard. Primer Command offers explainable insights and allows human operators to apply filters or geofences, ensuring critical events are surfaced for review while less relevant chatter is filtered out.
- **Dataminr** [11] provides a real-time AI platform that discovers the earliest signals of high-impact events, risks, and threats from public information streams. It continuously scans over a million public data sources worldwide, including social media, news sites, blogs, sensors, to detect anomalies and breaking events faster than any other source. Human analysts can review these AI-generated "briefs," trace the underlying sources, and decide on responses.
- **Signal AI** [12] offers an external intelligence platform for risk and reputation management, powered by its proprietary AI engine called AIQ. The system ingests 5+ million news articles, regulatory filings, and social

[10] `https://primer.ai/`

[11] `https://www.dataminr.com/`

[12] `https://signal-ai.com/`

media posts per day to "find the signal in the noise" for each client. A discriminative model first retrieves the most relevant data points, then a generative model produces concise insights which are grounded in the source material to minimise hallucinations. Humans supervise the process by tuning alert criteria, validating AI findings, and relying on audit trails that link back to original sources.

6 Related Work

A wide range of AI ethics guidelines and emerging regulations emphasise the need for meaningful human oversight of automated systems. For example, the EU's High-Level Expert Group on AI [10] identified "human agency and oversight" as a key requirement for trustworthy AI, and the proposed EU AI Act [11] explicitly mandates that high-risk AI systems be designed to allow effective human oversight. ISO/IEC 42001:2023 [12], the AI management system standard, emphasises oversight as governance routines and roles. The Australian Mandatory Guardrails for AI [13] also includes human oversight as a mandatory guardrail in high-risk settings, highlighting the need to minimise risks during the during the deployment and operation. While these regulations and standards articulate oversight at a high level, they provide limited actionable guidance on the concrete mechanisms and interaction contracts at the AI-human interface.

In academia, several studies focus on human oversight. Green [14] finds that oversight provisions in dozens of algorithmic governance policies often assume humans can reliably correct AI errors, an assumption undermined by evidence of human limitations. Studies have shown that humans sometimes reject correct AI outputs due to low trust or an inflated confidence in their own judgment, thereby nullifying potential benefits of the AI. Parasuraman and Manzey [15], for instance, describe how operators may either become complacent monitors who miss critical alerts, or conversely, intervene too often and "turn off" the automation even when it is performing well. Similarly, Green and Chen [16] observed that human decision-makers sometimes override accurate algorithmic recommendations, degrading overall decision quality.

Beyond generic "human-in-the-loop", philosophical accounts of Meaningful Human Control (MHC) argue for reason-responsive control distributed across design-time and run-time structures, such as who takes responsibility, what levers exist, and how systems track human values [17]. This complements regulatory oversight by clarifying the locus of control and responsibility in socio-technical systems, reinforcing our agency-first design stance.

Human-AI interaction (HAI) research examines what improves team performance in collaboration. As "explanations" alone rarely fix inappropriate reliance, interventions that change the interaction (e.g., who decides, when, and with what friction) tend to help more. Buçinca et al. [18] show that cognitive forcing functions (e.g., requiring an initial judgment before seeing AI) reduce over-reliance. Some studies report mixed or negative effects of generic transparency on users' ability to catch AI errors, reinforcing the need for task-tuned oversight affordances rather than one-size-fits-all Explanable AI (XAI). HAI design guidelines [19] consolidate such patterns into actionable rules (set expectations, support contestation, expose uncertainty, and adapt over time), which map well to our mechanisms catalogue.

In this study, existing regulations and standards explain why oversight is required; HAI and human-factors research suggest what tends to work; assurance research offers how to build it. We integrate these by reframing oversight responsibilities, shifting to external reasoning faithfulness, and providing a mechanisms-by-context catalogue of concrete oversight surfaces.

7 Conclusion

Human oversight should not be treated as a zero-sum game where adding human control necessarily diminishes AI autonomy. In this paper, we present a catalogue of oversight mechanisms and four end-to-end patterns to provide practical guidance for engineers and policymakers to operationalise oversight at both design time and run time. The mechanisms and patterns lay the foundation for AI systems that retain their useful autonomy while remaining aligned with human values and societal goals through informed, active human evaluation.

References

[1] Markus Langer, Veronika Lazar, and Kevin Baum. How to test for compliance with human oversight requirements in ai regulation? *arXiv preprint arXiv:2504.03300*, 2025.

[2] Johann Laux and Hannah Ruschemeier. Automation bias in the ai act: On the legal implications of attempting to de-bias human oversight of ai. *arXiv preprint arXiv:2502.10036*, 2025.

[3] David Abel, André Barreto, Michael Bowling, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, et al. Agency is frame-dependent. *arXiv preprint arXiv:2502.04403*, 2025.

[4] Xabier E Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive behavior*, 17(5):367–386, 2009.

[5] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

[6] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[7] Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. Measuring the faithfulness of thinking drafts in large reasoning models. *arXiv preprint arXiv:2505.13774*, 2025.

[8] Ramaravind K Mothilal, Joanna Roy, Syed Ishtiaque Ahmed, and Shion Guha. Human-aligned faithfulness in toxicity explanations of llms. *arXiv preprint arXiv:2506.19113*, 2025.

[9] United States Copyright Office. Copyright and artificial intelligence: Part 2—copyrightability. `https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf`, 2025. [Online; accessed 18-September-2025].

[10] European Commission High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, April 2019. [Online; accessed 18-September-2025].

[11] European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*, Jul. 12, 2024, July 2024. [Online; accessed 18-September-2025].

[12] ISO/IEC. ISO/IEC 42001:2023 Artificial intelligence — Management system. Geneva, Switzerland: International Organization for Standardization, 2023. [Standard].

[13] Australian Government, Department of Industry, Science and Resources. Mandatory ai guardrails. `https://www.industry.gov.au/publications/mandatory-ai-guardrails`, 2024. [Online; accessed 18-September-2025].

[14] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681, 2022.

[15] Raja Parasuraman and Dietrich H Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

[16] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019.

[17] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:323836, 2018.

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[19] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.