AI: Is a Human in the Loop a Sound Safeguard in Regulatory Settings?

A White Paper on Human-in-the-Loop Bias, Automation Overreliance, and Effective Oversight Design

Publication Information	Details
Publication Date	June 2026
Version	1.0
Authors	Eileen Williams, Certified AI Generalist; Lakshmi Tejaswi Sunkara, Certified AI Specialist
Contact	betweentheais@gmail.com
Website	www.betweentheais.com

Executive Summary

Regulators, courts, boards, and compliance teams increasingly rely on a familiar answer to artificial intelligence risk: keep a **human in the loop**. The concept is reassuring because it suggests that machine recommendations will remain subordinate to human judgment. Yet the safeguard is frequently treated as if human presence alone creates meaningful control. This white paper examines a narrower and more operationally important problem: **human in-the-loop bias**, especially the tendency of human reviewers to accept, approve, or insufficiently challenge AI or automated recommendations.

The central hypothesis is that in many regulated settings, humans are not functioning as independent safeguards. They are often **blindly following AI recommendations without**

question, particularly where systems appear authoritative, workflows are time-pressured, review tasks are repetitive, interfaces obscure uncertainty, or organizations implicitly reward speed and conformity. This phenomenon is commonly described as **automation bias**, overreliance, selective adherence, anchoring, or vigilance decrement. It can produce errors of commission, where a person follows a wrong automated recommendation, and errors of omission, where a person fails to act because an automated system did not alert them.[1]

This does not mean that human oversight is useless. Evidence from healthcare shows that human review can prevent serious harm when it is performed by trained domain experts who have real authority, sufficient time, clear escalation channels, and well-designed verification workflows. In one multicenter study, pharmacists intercepted 7,187 inpatient prescribing errors over six weeks; 46.6% had potentially serious consequences and 2.4% could have been life-threatening if not intercepted.[2] The lesson is therefore not that the human should be removed from regulated AI systems. The lesson is that **nominal human sign-off is not a sound safeguard** unless it is transformed into measurable, active, accountable, and independently auditable oversight.

The risks are not theoretical. Incidents in public benefits administration, employment screening, transportation automation, and healthcare demonstrate that human review can fail when it becomes passive, deferential, or institutionally constrained. Consequences include individual injury, denial of benefits, discrimination, reputational damage, regulatory action, civil settlements, and in safety-critical environments, potential loss of life. For organizations, the financial exposure includes direct settlement payments, remediation costs, litigation expenses, productivity losses, insurance implications, contract risk, and loss of revenue from public distrust or operational suspension.

Scope, Methodology, and Disclaimer

This white paper synthesizes publicly available research, regulatory materials, government reports, agency announcements, litigation-related public materials, peer-reviewed studies, and credible reporting available as of June 2026. The analysis is qualitative and policy-oriented. It is intended to identify recurring patterns, governance implications, and practical recommendations related to human-in-the-loop bias in regulated or high-impact settings.

> ***Standard disclaimer.*** *This publication is for general informational and educational purposes only. It does not constitute legal, medical, regulatory, technical, insurance, or professional advice. The authors have relied on public sources believed to be reliable,*

but no representation or warranty is made as to the completeness, accuracy, or timeliness of any source. Organizations should consult qualified legal, clinical, technical, risk, and compliance professionals before designing, deploying, auditing, or relying on AI systems in regulated settings.

1. The Regulatory Appeal of “Human in the Loop”

Human-in-the-loop governance is attractive because it appears to preserve human agency while allowing institutions to benefit from AI speed, scale, and predictive capacity. In regulated environments, the concept can support due process, professional accountability, safety review, discrimination prevention, and risk escalation. Laws and policy frameworks increasingly embed this intuition. For example, the EU AI Act requires high-risk AI systems to be designed so that they can be effectively overseen by natural persons, and it explicitly states that overseers should be aware of possible automatic reliance or overreliance on AI outputs.[3]

The legal and operational assumption is that a person reviewing a recommendation can notice error, bias, uncertainty, or contextual mismatch before harm occurs. In practice, however, human-in-the-loop designs often ask people to do the very thing humans are least suited to do: passively monitor a system that usually appears correct, then intervene instantly and decisively during rare failures. This is especially problematic when the AI system operates faster than human cognition, processes information the human cannot independently verify, or is embedded in a workflow where disagreement is costly.

Assumption in Nominal Oversight	Practical Failure Mode
A human reviewer will independently evaluate the AI output.	The reviewer anchors on the output and performs only superficial confirmation.
The reviewer can detect system error.	The model’s reasoning, data provenance, or uncertainty is not visible or understandable.
The reviewer has authority to override the AI.	Policies, productivity metrics, or organizational culture discourage override.
The reviewer has time to investigate anomalies.	Time pressure and workload turn review into checkbox compliance.

Human sign-off creates accountability.	Accountability shifts to the human while design, procurement, and governance failures remain underexamined.

The problem is therefore not simply “AI bias.” It is **human-in-the-loop bias**: the bias created by the interaction between machine recommendation, human psychology, workflow design, organizational incentives, and regulatory expectations.

2. Defining Human-in-the-Loop Bias

Human-in-the-loop bias occurs when a human reviewer’s judgment is systematically distorted by the presence, framing, authority, or workflow position of an AI or automated output. It is distinct from model bias, although the two often interact. A model may produce a biased, incomplete, or false recommendation; the human-in-the-loop problem arises when the reviewer fails to challenge it, selectively accepts it, or treats the system’s recommendation as a default truth.

Several related mechanisms matter in regulatory settings. **Automation bias** is the tendency to over-rely on automated systems, even where contradictory evidence exists.[1] **Overreliance** describes excessive trust in an AI output because the tool has been accurate before, appears sophisticated, or is institutionally endorsed. **Anchoring** occurs when the initial AI recommendation frames the reviewer’s subsequent thinking. **Selective adherence** occurs when human decision-makers accept algorithmic advice more readily when it aligns with existing stereotypes or prior beliefs.[4] **Vigilance decrement** describes
the decline in attention that occurs during prolonged passive monitoring, particularly where failures are rare.[5]

> *Human-in-the-loop bias is not the absence of a human. It is the presence of a human who has been placed in a workflow that makes independent judgment cognitively difficult, procedurally weak, or organizationally unrewarded.*

Recent empirical work supports this concern. In a 2026 experiment involving 2,784 participants reviewing AI-extracted values from corporate greenhouse gas emissions reports, researchers found that when correcting AI required additional effort, participants corrected fewer errors and accepted incorrect suggestions more often. Participants’ prior

attitudes toward AI were also strongly associated with performance: those more skeptical of AI detected errors more reliably, while those favorable toward automation accepted incorrect suggestions more often.6 This finding is directly relevant to regulatory settings because it shows that adding a human reviewer does not automatically neutralize AI error. The structure of the review task matters.

Healthcare studies show similar dynamics. A 2024 study of AI decision support defined automation bias as agreement with wrong AI-enabled diagnostic recommendations. It found that training, diagnostic performance, physician role, and other factors reduced false agreement, while perceived system benefit increased false agreement.7 The implication is uncomfortable but important: the people who may benefit from decision support may also be susceptible to overtrusting it unless training, workflow, and accountability are designed with that risk in mind.

3. Hypothesis: Humans Are Blindly Following AI Recommendations Without Question

The hypothesis of this white paper is that many organizations are overestimating the protective effect of human review. In practice, human reviewers may be approving AI or automated recommendations because they assume the system has already performed a superior analysis, because disagreement requires more documentation, because the interface encourages acceptance, or because the organization has not given them realistic authority to challenge the system.

The phrase “blindly following” should not be read as an accusation of carelessness by individual reviewers. In most cases, the deeper issue is systemic. Human reviewers may be competent and conscientious, yet still be placed in an impossible position. They may lack access to the data used by the model, lack time to perform an independent assessment, lack clarity about when override is expected, or fear consequences if they disagree with a tool that leadership has purchased, validated, or promoted.

Driver of Blind Acceptance	How It Appears in Practice	Regulatory Significance
Authority bias	The AI output is treated as more objective than human judgment.	Review may not satisfy meaningful oversight requirements.

Time pressure	Reviewers approve recommendations to meet throughput targets.	Compliance becomes procedural rather than substantive.
Opacity	Reviewers cannot see model limitations, confidence, or data lineage.	The human cannot evaluate what they cannot inspect.
Asymmetric documentation	Accepting the AI requires one click; overriding requires justification.	Workflow design nudges conformity.
Alert fatigue	Reviewers ignore warnings after repeated low-value alerts.	Safety signals lose operational value.
Organizational incentives	Employees learn that challenging automation slows work or creates conflict.	Governance failure is disguised as individual error.

In regulated settings, the danger is that human-in-the-loop becomes a legal fiction. The system appears compliant because a person clicked approve, signed a form, or remained available to intervene. Yet if that person was not realistically able to contest the recommendation, the safeguard may be more symbolic than substantive.

4. Risks: Adverse Outcomes, Loss of Life, and Revenue Loss

The risks from human-in-the-loop bias are multidimensional. They include direct human harms, legal and regulatory exposure, and business losses. The most severe risks arise in healthcare, transportation, public safety, infrastructure, financial services, employment, education, housing, insurance, and public benefits.

In safety-critical settings, a wrong recommendation accepted by a human reviewer can contribute to severe injury or loss of life. In administrative settings, erroneous automated decisions can deny benefits, employment, credit, housing, insurance, or services at scale. In commercial settings, failures can create settlement exposure, customer loss, contract termination, increased insurance premiums, and reputational decline.

Risk Category	Potential Harm	Example Consequence
Patient safety	Wrong diagnosis, incorrect medication, delayed care, missed deterioration.	Serious injury, preventable adverse drug event, malpractice claim, regulatory investigation.
Transportation and mobility	Passive supervision of automation fails during rare edge case.	Collision, injury, death, operational suspension, criminal or civil proceedings.
Employment and HR	Automated screening reinforces discriminatory exclusion.	EEOC action, settlement, monitoring obligations, reputational harm.
Public benefits and government services	Automated debt, eligibility, or enforcement decisions are accepted without meaningful contestability.	Large-scale unlawful collections, repayment, compensation, public inquiry, loss of trust.
Financial services and insurance	AI recommendations shape credit, fraud, pricing, or claims decisions.	Disparate impact claims, consumer protection enforcement, revenue loss, remediation costs.
Corporate governance	Human sign-off masks inadequate model risk management.	Board liability concerns, audit findings, regulatory penalties, contract risk.

Financial exposure is especially important because organizations often underestimate the revenue consequences of weak oversight. Settlement costs are only the visible portion. Robodebt in Australia involved repayment of more than A $751 million in unlawfully claimed debts and A $112 million in compensation to approximately 400,000 people.[8] In employment, iTutorGroup agreed to pay $365,000 to settle an EEOC discriminatory hiring suit involving alleged age discrimination through hiring software.[9] These figures do not capture internal investigation costs, reputational damage, leadership distraction, technology replacement, customer attrition, or the opportunity cost of delayed innovation.

5. Incidents to Date

The following incidents and studies illustrate how human-in-the-loop bias and weak oversight can appear across sectors. The purpose of this table is not to equate all cases or claim identical legal facts. Rather, it identifies recurring failure patterns: excessive trust in automated outputs, passive monitoring, inadequate contestability, poor escalation, and organizational reliance on human sign-off as proof of safety.

Incident or Evidence Base	Sector	Human-in-the-Loop Bias Pattern	Harm or Exposure
Uber ATG Tempe fatal crash	Autonomous vehicle testing	A safety driver was expected to monitor a developmental automated driving system during public-road testing, but passive supervision failed during a rare edge case. The NTSB emphasized inadequate safety culture and safety risk management.10	A pedestrian was fatally injured. The case illustrates that human presence may not compensate for automation design, delayed handoff, vigilance decrement, and weak safety governance.
Robodebt	Public benefits administration	Automated income averaging generated inaccurate debts, and institutional processes placed the burden on individuals to disprove debts rather than ensuring meaningful human review before enforcement.8	More than half a million debts were raised; the settlement involved repayment and compensation to approximately 400,000 people.8
iTutorGroup EEOC settlement	Employment screening	Hiring software allegedly screened out older applicants. Any human review process failed to prevent or remediate the discriminatory effect before enforcement action.9	The company agreed to pay $365,000 to settle the EEOC suit.9

AI-assisted diagnostic decision support studies	Healthcare	Participants or clinicians may agree with incorrect AI recommendations, especially where the system is perceived as beneficial or authoritative.7	Wrong clinical decisions can lead to delayed treatment, misdiagnosis, or patient harm if not caught.
AI-assisted data extraction experiment	Compliance and reporting	Participants accepted incorrect AI-generated suggestions more often when correction required greater effort, showing that review friction affects error detection.6	In regulated reporting, this can create inaccurate disclosures, audit risk, and compliance exposure.

Incident or Evidence Base	Sector	Human-in-the-Loop Bias Pattern	Harm or Exposure
Computerized order entry hard-stop warning	Healthcare technology	A strong automated intervention reduced risky co-prescribing but created treatment delays, leading the trial to stop and the warning to be abandoned.11	The case shows that both blind reliance and blunt blocking can harm patients if not monitored and tuned.

The Uber ATG crash is particularly important because it shows the limits of passive human supervision. The NTSB report states that a developmental automated driving system was active when the vehicle struck and fatally injured a pedestrian.10 Public reporting on NTSB documents described repeated misclassification of the pedestrian, delayed braking behavior, and a late handoff to the safety driver.12 A human was technically in the loop, but the design required sustained vigilance over automation until the moment of failure. This is a weak human-in-the-loop control.

Robodebt illustrates a different version of the same problem. The danger was not a split second failure but a mass administrative process in which automated or semi-automated

outputs were institutionally treated as enforceable unless affected individuals disproved them. Where the human review function is procedural, constrained, or displaced onto the public, oversight does not meaningfully protect rights.

6. Where Human Oversight Worked: A Patient-Safety Counterexample

A balanced assessment must acknowledge that human oversight can prevent harm. The strongest safeguards are not passive sign-offs. They are active verification practices performed by trained professionals at well-designed control points.

Cabri and colleagues studied pharmacist interception of prescribing errors across 11 hospitals. Pharmacists documented 7,187 intercepted prescribing errors during inpatient order verification over six weeks. Of those errors, 3,349 were assessed as potentially serious and 175 as potentially life-threatening if not intercepted.2 This is a powerful example of human-in-the-loop oversight working in a patient-safety context. It shows that trained humans can catch errors and prevent severe harm when the workflow gives them a clear role, relevant clinical context, standardized documentation, and authority to intervene before an order reaches the patient.

Another patient-safety example involves computerized provider order entry and wrong patient orders. Green and colleagues evaluated a patient verification dialog in five emergency departments. The intervention required clinicians to confirm patient identity after a mandatory 2.5-second delay, and wrong-patient orders fell by 30% immediately after implementation; after two years, the rate remained 24.8% lower than before the intervention.13 This study is relevant because it did not merely place a human somewhere in the process. It changed the interaction design so the human had to perform a specific verification act at the right moment.

Why These Examples Worked	Practical Meaning for AI Oversight
The human was a trained professional.	Oversight should be assigned to qualified domain experts, not generic approvers.
The review occurred before harm reached the patient.	Human review must be placed upstream of irreversible or high-impact action.
The workflow required specific verification.	Oversight should require targeted checks, not a general approval click.

The intervention was documented.	Override, acceptance, and rationale should be auditable.
The system supported human judgment.	Interfaces should reveal relevant context and make challenge operationally feasible.

These cases should inform regulation. A human-in-the-loop safeguard is sound only when it resembles pharmacist verification or patient-identity confirmation: active, informed, documented, and empowered. It is not sound when it resembles passive monitoring, rushed approval, or after-the-fact blame assignment.

7. Impacted and Interested Parties

Human-in-the-loop bias affects a broad ecosystem. The most obvious impacted parties are individuals who receive automated or AI-influenced decisions. However, the interested party landscape is wider because responsibility is distributed across developers, deployers, procurement teams, compliance officers, boards, insurers, regulators, and the public.

Party	Interest or Exposure
Patients, consumers, employees, applicants, beneficiaries, and citizens	Risk of incorrect, discriminatory, unsafe, or unchallengeable decisions.
Clinicians, caseworkers, drivers, analysts, and frontline reviewers	Risk of becoming the “liability sponge” for system failures they could not realistically detect or prevent.[5]
AI developers and vendors	Exposure for inadequate design, insufficient validation, misleading performance claims, weak explainability, or poor monitoring support.
Deploying organizations	Duty to implement meaningful oversight, train staff, monitor performance, investigate failures, and maintain audit trails.
Boards and senior leadership	Governance responsibility for risk appetite, AI controls, escalation, and public accountability.

Regulators and enforcement agencies	Need to distinguish nominal human review from effective oversight and to define evidence of compliance.
Insurers and risk managers	Need to price and manage AI-related operational, professional liability, cyber, and directors-and-officers risks.
Auditors and compliance professionals	Need to evaluate model governance, human review quality, override rates, error rates, and process integrity.
Civil society and the public	Interest in transparency, contestability, fairness, safety, and institutional trust.

The frontline human deserves special attention. Many AI governance frameworks implicitly place responsibility on the reviewer who accepts or rejects a recommendation. Yet if that person lacks meaningful capacity to review, the organization has created a **moral crumple zone**: the human absorbs blame for failures created by system design, procurement, staffing, incentives, or policy.[5]

8. Recommendations to Rectify

The solution is not to abandon human oversight. The solution is to redesign it. Regulators and organizations should replace the generic phrase “human in the loop” with specific, testable oversight requirements. The question should not be, “Was a human present?” The question should be, “Could the human reasonably detect, challenge, document, and stop the AI-driven action before harm occurred?”

8.1 Define Oversight by Function, Not Presence

Regulatory and organizational policies should define the oversight function in operational terms. The policy should specify who reviews the output, what information they must inspect, what authority they have, when they must escalate, what documentation is required, and how the organization measures review quality. A human-in-the-loop safeguard should not be credited unless the reviewer has the knowledge, time, information, and authority to intervene.

8.2 Design for Productive Friction

Many workflows make acceptance easy and disagreement hard. This is dangerous. Organizations should introduce **productive friction** for high-impact decisions, such as requiring second review, uncertainty review, evidence inspection, or supervisor escalation when model confidence is low, protected-class risk is elevated, or the decision has irreversible consequences. Friction should be risk-tiered so that it improves safety without producing dangerous delays, as the healthcare hard-stop warning case demonstrates.[11]

Oversight Design Element	Weak Version	Strong Version
Review action	One-click approval.	Structured review with required evidence checks.
Override process	Override requires burdensome justification; acceptance does not.	Acceptance and override both require proportionate rationale in high-risk cases.
Uncertainty display	Single recommendation without confidence or limits.	Confidence, known limitations, data gaps, and alternative explanations are visible.
Escalation	Reviewer must decide alone.	Defined escalation pathways for uncertainty, disagreement, or severe impact.
Auditability	Logs show only final approval.	Logs capture recommendation, human action, rationale, override, and subsequent outcome.

8.3 Measure Overreliance and Override Quality

Organizations should track whether human reviewers are actually exercising judgment. Relevant metrics include acceptance rates, override rates, override outcomes, time spent reviewing, disagreement patterns by reviewer, error discovery rates, appeal outcomes, demographic impact, and post-decision harm indicators. A near-100% acceptance rate may indicate that the AI is excellent, but it may also indicate rubber-stamping. The difference can only be known through audit.

8.4 Train Reviewers on Human-in-the-Loop Bias

Training should explicitly address automation bias, anchoring, selective adherence, alert fatigue, and vigilance decrement. Reviewers should practice identifying wrong recommendations, not merely learn how to use the tool. Training should include adversarial examples, uncertainty interpretation, documentation expectations, escalation procedures, and accountability boundaries.

8.5 Separate Model Validation from Operational Oversight

A frontline reviewer should not be the first or only defense against model error. Organizations should conduct pre-deployment validation, subgroup performance testing, red-team review, usability testing, workflow simulation, and post-market monitoring. Human review is one layer of defense, not a substitute for model governance.

8.6 Protect the Right to Challenge and Appeal

For decisions affecting rights, benefits, employment, housing, healthcare, credit, insurance, education, or liberty interests, affected individuals should receive meaningful notice and a practical way to contest the decision. Contestability is an external safeguard against internal overreliance. Robodebt demonstrates the danger of systems that shift the burden to individuals without adequate pre-enforcement review.[8]

8.7 Create Independent Escalation Channels

Reviewers should be able to raise concerns about AI recommendations without retaliation or productivity penalty. This includes whistleblower-style escalation for suspected systemic error, bias, data drift, unsafe recommendations, or pressure to accept outputs. Serious concerns should trigger incident review and, where appropriate, temporary suspension of the system.

8.8 Align Procurement and Contracts with Oversight Duties

Contracts with AI vendors should require documentation, performance evidence, audit rights, incident reporting, data lineage, limitations disclosure, human factors testing, and post-deployment monitoring support. A deployer cannot create effective human oversight if the vendor does not provide enough information to evaluate system behavior.

8.9 Use Two-Person or Independent Review for Severe-Risk Decisions

Where decisions carry a risk of severe injury, deprivation of essential services, major financial loss, or irreversible legal consequence, organizations should consider two-person verification or independent review. The EU AI Act's special treatment of certain high-risk systems reflects the broader principle that some decisions require more than one reviewer and more than ordinary awareness of overreliance.[3]

8.10 Treat Human-in-the-Loop Failure as an AI Incident

When a human accepts an erroneous AI recommendation, the event should not be categorized only as human error. It should trigger an AI incident review that examines model output, interface design, workload, training, policy, incentives, alerting, documentation, and governance. This prevents organizations from blaming the last human in the chain while leaving systemic causes intact.

9. Practical Oversight Maturity Model

Organizations can use the following maturity model to assess whether their human-in-the loop approach is symbolic or substantive.

Maturity Level	Description	Risk Posture
Level 1: Nominal Sign-Off	A human approves AI output with little guidance, evidence, or documentation.	High risk of rubber stamping and liability transfer.
Level 2: Procedural Review	Review steps exist, but reviewers lack training, time, or meaningful override authority.	Moderate to high risk; compliance may be superficial.
Level 3: Structured Oversight	Reviewers are trained, required checks are documented, and overrides are tracked.	Improved control; requires monitoring for overreliance.
Level 4: Risk-Tiered Oversight	Oversight intensity changes based on impact, uncertainty, protected-class risk, and reversibility.	Stronger alignment with regulated-risk expectations.

Level 5: Audited Human-AI Governance	Human review quality, AI performance, appeals, incidents, and outcomes are continuously audited and improved.	Best practice for high impact and safety-critical systems.

The goal for regulated settings should be Level 4 or Level 5. Anything below Level 3 should be treated as insufficient for high-risk decisions.

10. Conclusion

A human in the loop can be a sound safeguard in regulatory settings, but only under specific conditions. It is not sound when it means passive monitoring, rushed approval, unclear authority, or symbolic sign-off. Human-in-the-loop bias makes this distinction urgent. The presence of a human may create confidence for regulators and the public while leaving the actual decision process vulnerable to automation bias, overreliance, anchoring, alert fatigue, and organizational pressure.

The incidents and studies reviewed in this white paper support a practical conclusion: **human oversight must be designed, trained, measured, and audited**. It must be treated as a control system, not a slogan. The most effective examples, such as pharmacist interception of potentially life-threatening prescribing errors, show that human oversight can prevent severe harm when it is active, expert, and embedded at the right control point.2 The most troubling examples show that nominal human presence can fail catastrophically when the task is cognitively unrealistic, procedurally weak, or organizationally unsupported.10

For regulated AI, the standard should shift from “human in the loop” to **human with the capacity, authority, and obligation to challenge the loop**. Organizations that make this shift will be better positioned to protect people, reduce legal exposure, preserve revenue, and build durable trust in AI-enabled systems.

References

Between the AIs
www.betweentheais.com
betweentheais@gmail.com

Publication Date: June 2026 | Version 1.0