Judgment Decay in AI-Augmented Environments

A Framework for Continuous Measurement and Embedded Maintenance of Human Decision-Making Capability

Peyton Flock

COHESION Research, Spokane, WA

peyton@cohesionauth.com

April 13, 2026

Preprint — Open Access

Abstract

AI decision-support systems are increasingly assuming cognitive tasks historically central to professional expertise across healthcare, aviation, criminal justice, and finance. While automation complacency has been studied for decades (Parasuraman & Manzey, 2010), and the fundamental paradox of automation was identified over forty years ago (Bainbridge, 1983), no integrated framework exists for continuously measuring, actively maintaining, and verifiably documenting the progressive deterioration of independent decision-making capability as it occurs across AI-augmented professional environments. This paper introduces the construct of Judgment Decay to describe this phenomenon — the progressive, measurable deterioration of independent decision-making capability that occurs when professionals routinely defer to AI recommendations.

I build on complacency research from aviation and automation (Parasuraman & Manzey, 2010; Ebbatson et al., 2010), self-efficacy and autonomy theory from clinical psychology (Bandura, 1977; Deci & Ryan, 2000), situation awareness theory (Endsley, 1995), and dual-process cognition (Kahneman, 2011) to propose the Judgment Maintenance Framework — an integrated architecture with four components: (1) continuous measurement of judgment decay through AI interaction patterns, (2) invisible maintenance through clinically informed micro-interventions embedded within AI systems, (3) continuous verification via a composite Judgment Independence Score, and (4) embedded integration at the AI system level. I argue that effective judgment maintenance requires architectural integration within AI systems themselves — imperceptible to the end user, quantifiable by the deploying organization, and auditable by regulators. The framework directly addresses the EU AI Act's mandate for "effective human

oversight" (Article 14) by operationalizing a requirement that the regulation mandates but does not define.

This paper makes three contributions. First, it introduces Judgment Decay as a formally defined construct distinct from automation complacency, automation bias, and skill atrophy, integrating these previously separate literatures under a unified clinical framework. Second, it proposes the Judgment Maintenance Framework grounded in self-efficacy theory, self-determination theory, and desirable difficulties research. Third, it operationalizes the EU AI Act's oversight mandate by providing measurable indicators and a composite Judgment Independence Score.

Keywords: judgment decay, AI safety, human oversight, automation complacency, EU AI Act, self-efficacy, decision-making, human factors

1. Introduction: The Structural Erosion of Human Judgment Under AI Augmentation

A significant shift is occurring in the nature of professional expertise, one that has received insufficient attention in the human factors, AI safety, and organizational psychology literatures.

AI decision-support is different from every previous form of automation. Manufacturing robots displaced physical labor. Spreadsheets displaced calculation. But AI-powered recommendation systems displace something more fundamental: the cognitive work that professionals identify as their expertise. Pattern recognition under uncertainty. Contextual overrides of formal rules. The integration of years of embodied experience into real-time judgment calls.

The upside is documented extensively — speed, fewer routine errors, specialist-level analysis available to generalists. What remains insufficiently documented — in part because no standardized measurement framework exists — is the corresponding cost to human capability.

The human capabilities that AI systems are designed to augment are the same capabilities that atrophy when AI systems do the augmenting.

Applied fieldwork in tacit knowledge extraction with retiring manufacturing professionals provided the experiential foundation for this paper. Direct observation of how professional judgment is maintained, transferred, and lost under conditions of workforce transition revealed something the automation complacency literature describes abstractly but rarely captures at the individual level: the relationship between practiced judgment and professional identity is not metaphorical. When people stop exercising the skills that define their expertise, they do not simply lose capability. They lose a piece of who they are. That observation — clinical in origin, applied in context — is the foundation of this framework.

1.1 The Mechanism

Judgment is not knowledge. It is not information retrieval. It is a practiced skill — one that develops through exercise and deteriorates without it. This distinction matters enormously because it means judgment cannot be preserved by documentation, refresher courses, or periodic testing alone. Like any practiced skill, it requires ongoing use. Motor skill research has demonstrated this for decades (Arthur et al., 1998), situation awareness degrades without active exercise of monitoring capabilities (Endsley, 1995), and aviation research has shown measurable correlations between recent manual flying experience and handling performance (Ebbatson et al., 2010). Expertise research more broadly establishes that high-level performance requires sustained deliberate practice (Ericsson & Charness, 1994).

When an AI system provides a recommendation, the human operator faces a choice: evaluate independently and compare, or defer to the AI. Research on automation bias demonstrates that operators overwhelmingly defer (Parasuraman & Manzey, 2010; Skitka et al.,

2000), particularly when the AI system has a high stated confidence level, the operator is under time pressure, the AI has been historically accurate, or the cost of independent evaluation exceeds the perceived cost of deference. In dual-process terms, deference is a System 1 heuristic response — fast, effortless, and locally rational (Kahneman, 2011). Independent evaluation requires the effortful engagement of System 2 processing.

Critically, each individual act of deference is locally rational. The AI system is typically faster and more consistent; the cost of independent evaluation is nontrivial; and the operator's prior experience confirms that AI recommendations are usually correct. Deference under these conditions is predictable and well-documented (Parasuraman & Manzey, 2010).

But compounded across thousands of decisions, something shifts. An operator who defers on 90% of decisions may, over subsequent months, show increasing deference rates — not because the AI system improved, but because the operator's independent judgment capability degraded through disuse. While this specific trajectory has not been measured (because no measurement framework exists), the underlying mechanism — skill decay from non-practice — is one of the most robust findings in human performance research (Arthur et al., 1998).

1.2 The Scale

Preliminary evidence from domains with advanced AI adoption suggests that these dynamics are already observable:

Aviation: Pilots who had flown more sectors in the preceding week demonstrated measurably better heading and yaw control performance (Ebbatson et al., 2010), suggesting that even short periods of reduced manual practice produce detectable skill degradation. The FAA has issued guidance (SAFO 17007) encouraging airlines to ensure pilots periodically hand-fly aircraft, and a 2022 report recommended that airline training programs "promote the development and

maintenance of" manual flying skills (FAA, 2022) — an implicit acknowledgment that automation degrades the skills it was designed to augment. NHTSA's investigation documented 956 crashes with Tesla Autopilot engaged between 2018 and 2023, including 29 fatalities, finding that "foreseeable driver misuse of the system played an apparent role" and that "Tesla's weak driver engagement system was not appropriate for Autopilot's permissive operating capabilities" (NHTSA, 2024). The phenomenon of automation surprises — moments when automated systems behave unexpectedly and operators cannot respond effectively — has been well-documented as a direct consequence of skill degradation under automation (Sarter et al., 1997).

Healthcare: AI adoption in diagnostic imaging has expanded rapidly, with over 900 FDA-cleared AI algorithms available as of 2024 (ACR AI Central), and major health systems deploying AI across imaging workflows. AI diagnostic systems have achieved accuracy exceeding 90% for specific imaging tasks in controlled settings (McKinney et al., 2020), which raises a critical question: if the AI handles the majority of routine cases correctly, how does the radiologist maintain competence on the cases the AI misses? Automation bias research in healthcare suggests that clinicians are not immune to over-reliance on automated recommendations (Goddard et al., 2012). Recent empirical work with 319 knowledge workers found self-reported reductions in critical thinking effort when using generative AI — direct evidence that AI use may degrade the cognitive engagement that professional judgment requires (Lee et al., 2025). To date, no standardized framework exists for measuring this form of capability degradation longitudinally.

Criminal Justice: COMPAS risk assessment tools are used in Wisconsin, Florida, and New York for bail and sentencing decisions. ProPublica's analysis demonstrated racially asymmetric

error rates — Black defendants who did not reoffend were nearly twice as likely to be classified as high-risk compared to white counterparts (Angwin et al., 2016). Subsequent research has shown that certain fairness criteria are mathematically incompatible when base rates differ across groups (Chouldechova, 2017), deepening concerns about algorithmic decision-support in criminal justice. France has banned the use of AI to analyze or predict the behavior of individual judges (Article 33, Justice Reform Act, 2019). Judges may increasingly defer to algorithmic risk scores, potentially reducing the exercise of independent judicial reasoning — a pattern consistent with the judgment decay framework proposed here, with significant implications for due process.

Finance: Algorithmic trading systems execute decisions at speeds that preclude meaningful human oversight. A ProPublica investigation found that Cigna's automated PxDx system enabled physicians to deny over 300,000 claims in two months, spending an average of 1.2 seconds per review (Angwin et al., 2023) — a rate that precludes meaningful medical judgment on individual cases. When these systems produce systematic errors, the human operators tasked with oversight may have lost the domain expertise necessary to detect them.

1.3 The Paradox

Bainbridge (1983) identified what she termed the "ironies of automation" — the observation that automating a task creates new human performance problems rather than eliminating them. The defining paradox of AI-augmented work extends this insight: **the more reliable the AI system, the faster human judgment decays, and the more catastrophic the consequences when the AI fails.**

A system that is correct 99% of the time creates operators who exercise judgment 1% of the time. When the system encounters a situation outside its training distribution — which is

inevitable in any complex domain — the human who is supposed to catch the failure has not exercised the relevant judgment capability in months or years.

Documented AI safety incidents increased 56.4% from 149 in 2023 to 233 in 2024 (Stanford HAI AI Index Report, 2025). In 2025, 40+ researchers from Anthropic, Google DeepMind, OpenAI, and Meta co-authored a joint warning that there is "no guarantee that the current degree of visibility" into AI reasoning "will persist" as models advance (Lanham et al., 2025), with research showing reasoning models may obscure their true decision processes. AI systems are becoming less interpretable at the same time that the human operators responsible for oversight may be losing the interpretive capabilities required to perform that function.

This paper proposes the Judgment Maintenance Framework, an integrated architecture for the continuous measurement, invisible maintenance, and regulatory verification of independent human judgment in AI-augmented professional environments. I pose two research questions:

RQ1: Can independent judgment capability be validly measured through patterns of human-AI interaction without separate assessment instruments?

RQ2: Can clinically informed micro-interventions, embedded invisibly within AI systems, maintain judgment capability that would otherwise decay?

The remainder of this paper is organized as follows. Section 2 reviews the theoretical foundations underlying the framework. Section 3 examines why existing approaches to skill maintenance are insufficient. Section 4 presents the four-component Judgment Maintenance Framework with testable propositions. Section 5 maps the framework to EU AI Act requirements. Section 6 describes the proposed implementation architecture and its invisible

design constraint. Section 7 addresses ethical considerations. Section 8 discusses implications and future directions. Section 9 acknowledges limitations, and Section 10 concludes.

2. Theoretical Background

The Judgment Decay construct integrates four previously separate research traditions. This section reviews each and identifies the specific gap the present framework addresses.

2.1 Automation Complacency and Automation Bias

The automation complacency literature, anchored by Parasuraman and Manzey's (2010) attentional integration model, establishes that human monitoring performance degrades as a function of automation reliability and monitoring duration. Operators exhibit vigilance decrements within 30 minutes of monitoring highly reliable automated systems. Skitka, Mosier, and Burdick (2000) provided seminal empirical evidence of automation bias — the tendency to follow automated recommendations even when contradictory information is available — demonstrating both omission errors (failing to act when automation fails to alert) and commission errors (acting on incorrect automated recommendations).

Bainbridge (1983) identified the fundamental paradox: automation designed to reduce human error introduces new categories of human error by degrading the skills and situation awareness needed to intervene when automation fails. This insight, articulated over four decades ago, has only become more relevant as AI systems assume cognitive rather than merely procedural functions.

2.2 Skill Decay and Expertise Maintenance

Arthur, Bennett, Stanush, and McNelly's (1998) meta-analysis of skill decay established that skills deteriorate significantly with nonuse, with effect sizes exceeding d = -1.4 after periods of

one year or more. The rate of decay varies by skill type, with cognitive and procedural skills showing different degradation curves. Ericsson and Charness (1994) established that expert performance requires sustained deliberate practice — a finding that implies expertise cannot be maintained passively even when knowledge is retained.

In aviation, the relationship between recent manual flying experience and handling performance is well-documented (Ebbatson et al., 2010), and regulatory bodies have responded with guidance encouraging periodic hand-flying (FAA SAFO 17007). This represents the closest existing precedent to the judgment maintenance approach proposed here.

2.3 Situation Awareness and Cognitive Engagement

Endsley's (1995) three-level model of situation awareness — perception, comprehension, and projection — maps directly onto the capabilities at risk under judgment decay. AI systems that provide pre-processed recommendations may degrade all three levels: operators perceive less raw data (the AI pre-filters), comprehend fewer relationships (the AI synthesizes), and project fewer future states (the AI predicts). Wickens, Hollands, Banbury, and Parasuraman (2021) extend this through Multiple Resource Theory, explaining why operators cannot simultaneously maintain full situation awareness while deferring cognitive processing to AI — the same attentional resources are required for both.

2.4 Behavioral Intervention and Choice Architecture

Thaler and Sunstein (2008) established that environmental modifications to decision architecture — nudges — can produce significant behavioral change without restricting choice. Sweller's (1988) cognitive load theory explains that appropriate levels of cognitive engagement are necessary for skill maintenance, providing the theoretical basis for the micro-intervention approach proposed here. Bjork and Bjork's (2011) desirable difficulties framework demonstrates

that introducing calibrated challenges during learning and practice enhances long-term retention — a principle the present framework applies to operational contexts.

2.5 The Gap

The National Academies of Sciences (2022) identified situation awareness, appropriate trust calibration, and training for human-AI teams as critical research needs. What is missing from the existing literature is an integrated framework that: (a) continuously measures judgment capability degradation as it occurs within operational AI interactions, (b) embeds maintenance interventions within AI systems at the architectural level, and (c) connects measurement and maintenance to regulatory compliance requirements. The Judgment Maintenance Framework proposed in Section 4 addresses this gap.

3. Why Existing Approaches Are Insufficient

The conventional institutional response to capability gaps is retraining. This section argues that retraining and related approaches are insufficient for the specific problem of judgment decay, and that understanding their limitations clarifies the design requirements for an effective alternative.

3.1 Training Programs

Judgment develops within the operational environment — under real time pressure, with real consequences, embedded in the specific decision architecture of a particular role. Training programs extract the learner from that context, place them in a simulated one, and hope the skills transfer back. Training transfer research demonstrates that the gap between learning in training and performance in operations is substantial, particularly when the operational environment differs from training conditions (Baldwin & Ford, 1988). For complex judgment tasks embedded

in specific decision architectures, this transfer gap is likely wider still. The deeper problem is temporal: training happens once or twice a year; judgment decay may occur with every AI-assisted decision between sessions.

3.2 Credential-Based Verification

Professional credentials verify a snapshot. They confirm that at some point in the past, this person demonstrated sufficient knowledge to pass an examination. They say nothing about whether that person can still exercise the judgment their credential implies. Consider two board-certified radiologists — one who independently evaluates every scan, one who has deferred to AI for two years. Their credentials are identical. Their capabilities are not. No existing system distinguishes between them.

Credentials are even worse as a solution when the relevant capability is not knowledge recall but practiced judgment — something that no examination reliably measures.

3.3 Monitoring and Separate Assessment

External monitoring of human-AI interaction (e.g., logging how often operators override AI recommendations) captures behavior but not capability. An operator who never overrides the AI might have excellent judgment and agree with correct recommendations, or might have severely decayed judgment and be rubber-stamping everything. Behavior alone cannot distinguish these states.

Any assessment that exists outside the operational workflow — a test, a survey, a simulation — suffers from Hawthorne effects. Operators perform differently when they know they are being assessed. Judgment decay may be particularly insidious because it is, by its nature, invisible to the individual experiencing it.

4. The Judgment Maintenance Framework

The framework has four parts. All of them are designed to disappear — the user should never realize they are being maintained, never feel friction, never have their workflow interrupted. If the operator notices it, it has failed.

4.1 Component 1: Continuous Measurement Through Interaction Patterns

Principle: The pattern of how a human interacts with an AI system is hypothesized to contain sufficient information to estimate the state of their independent judgment capability, reducing or eliminating the need for separate assessment. This hypothesis requires empirical validation; the indicators proposed below are theoretically motivated proxies whose construct validity must be established through controlled studies comparing interaction-derived scores against independent performance measures.

Measured indicators:

Indicator	What It Reveals	Collection Method
Time-to-acceptance	Speed of deference to AI recommendation	Timestamp analysis
Modification frequency	How often the human alters AI output	Edit tracking
Override rate	How often the human rejects AI recommendation	Decision logging
Independent performance delta	Quality difference when AI is unavailable vs. available	Periodic withholding
Error detection rate	Ability to catch incorrect AI output	Calibrated error injection
Deliberation depth	Engagement with alternatives before deciding	Interaction pattern analysis
Post-error behavior change	Whether encountering an AI error changes subsequent behavior	Longitudinal tracking

Clinical basis: This approach extends Parasuraman and Manzey's (2010) automation complacency framework, which established that operator vigilance degrades as a function of automation reliability and monitoring duration. The present framework operationalizes this finding as a continuous measurement protocol. Endsley's (1995) situation awareness model

provides the theoretical structure for interpreting these indicators as proxies for SA degradation at the perception, comprehension, and projection levels.

A critical construct validity concern: several of these indicators are ambiguous in isolation. Rapid time-to-acceptance may reflect either high expertise (the operator quickly confirms a correct recommendation) or advanced decay (the operator rubber-stamps without evaluation). The framework addresses this through composite scoring — no single indicator is diagnostic. The Independent Performance Delta, measured through periodic AI withholding, provides the ground-truth anchor against which other indicators are calibrated. Without this anchor, the remaining indicators would be insufficient.

4.2 Component 2: Invisible Maintenance Through Micro-Interventions

Principle: When measurement detects judgment decay, the AI system itself adjusts its behavior to exercise the human's judgment — without the human knowing the adjustment has occurred.

Withholding: The AI periodically withholds its recommendation and presents the decision to the human without assistance. The human perceives this as the AI "still processing" or "needing additional input." The frequency of withholding is calibrated to the operator's current judgment score — more frequent for operators showing decay, less for those maintaining capability.

Clinical basis: Mastery Experience (Bandura, 1977). Self-efficacy is built through successful independent performance. A potential limitation: Bandura's model requires that the individual attribute success to their own capability, and the invisible design constraint means the operator may not consciously register the experience as independent performance. I hypothesize that the skill maintenance effect (preserving practiced capability through exercise) operates independently of the self-efficacy effect (building confidence through attributed success). The

withholding intervention targets the former; the self-efficacy benefit, while likely reduced by non-attribution, is a secondary gain. Empirical work is needed to determine whether covert forced independence produces comparable skill maintenance to overt independent practice.

Split presenting: Instead of a single recommendation, the AI presents two or more equally viable options without indicating preference. The human must evaluate and choose, exercising evaluative judgment — which prior research suggests is among the earliest capabilities to degrade under conditions of automation complacency (Parasuraman & Manzey, 2010). *Clinical basis:* Self-Determination Theory, Autonomy dimension (Deci & Ryan, 2000). Presenting genuine choices preserves the experience of autonomous decision-making, which is intrinsically motivating and maintains cognitive engagement. The frequency of split-presenting must be calibrated against the decision fatigue literature (Baumeister et al., 1998) — excessive forced choice under cognitive load can degrade rather than maintain performance. The adaptive difficulty scaling described below addresses this constraint by reducing intervention frequency when the operator demonstrates maintained capability.

The third mechanism is more aggressive. **Calibrated challenge** involves the AI occasionally introducing content that contains a subtle error or inconsistency, designed to be detectable by a competent professional but not obvious to a disengaged one. The human's response — detection versus miss — is the highest-fidelity measure of judgment quality available. This extends the concept of "desirable difficulties" in learning (Bjork & Bjork, 2011), which was validated in educational settings. Critical constraint: in safety-critical domains (healthcare, aviation, infrastructure), calibrated challenges must operate exclusively in sandboxed decision contexts where the outcome cannot propagate to real-world consequences — for example, during case review, training queues, or decision audit workflows rather than live

patient care or active flight operations. The system must include a failsafe: if the operator fails to detect an injected error, the system intercepts the decision before execution. The application of desirable difficulties to operational environments with real consequences requires domain-specific safety protocols that this framework identifies as necessary but does not specify.

These interventions are governed by **adaptive difficulty scaling**: as the operator demonstrates maintained judgment, intervention frequency decreases; as decay is detected, frequency increases. The system operates as an adaptive maintenance environment that the operator never perceives as such. *Clinical basis:* Zone of Proximal Development (Vygotsky, 1978) and flow state theory (Csikszentmihalyi, 1990) — optimal cognitive engagement occurs when challenge is calibrated to capability.

An additional theoretical contribution connects Acceptance and Commitment Therapy (Hayes et al., 2006) to the deference mechanism. ACT's concept of experiential avoidance — the tendency to avoid aversive internal experiences — maps onto the operator's avoidance of cognitive effort when deferring to AI. Independent evaluation is cognitively demanding and carries the risk of discovering one's own errors; deference avoids both. The withholding intervention can be understood as interrupting an avoidance pattern, forcing contact with the aversive experience (uncertainty, cognitive effort) that the operator has learned to bypass through AI deference. This connection between experiential avoidance and automation complacency has not, to my knowledge, been previously drawn in the literature.

4.3 Component 3: Continuous Verification

Principle: Human verification should not be a point-in-time check. It should be a continuous, passive signal derived from real interaction behavior.

Current verification systems (CAPTCHA, biometric scanning, credential checks) answer a binary question: "Is this a human?" The relevant question for AI-augmented environments is: "Is this human still exercising judgment?"

The interaction pattern itself is the verification. An operator who regularly modifies AI output, detects injected errors, and demonstrates appropriate skepticism toward high-confidence recommendations would score high on the proposed Judgment Independence metric. An operator who consistently accepts output without modification, fails to detect calibration errors, and shows no behavioral variation as a function of AI confidence would score low — suggesting possible judgment decay.

This continuous verification generates a **Judgment Independence Score** — a composite metric derived from all measured indicators, updated with every interaction, and trackable over time. This score provides organizations and regulators with the first quantified measure of whether human oversight is real or performative.

4.4 Component 4: Embedded Integration

Principle: Judgment maintenance cannot function as a separate product. It must be embedded within the AI systems people already use.

This is the foundational design constraint. If judgment maintenance requires the operator to use a separate tool, take a separate test, or engage in a separate activity, adoption will be limited to mandated contexts and will not reach the scale necessary to address the problem.

The framework is designed for implementation as middleware — a layer between the AI model and the user interface that intercepts, measures, and occasionally modifies AI system behavior. From the user's perspective, nothing changes. From the AI system's perspective, an

additional processing layer has been inserted. From the organization's perspective, a compliance and measurement dashboard has been activated.

The closest public health analogy — imperfect but instructive — is the fortification of staple foods (iodine in salt, folic acid in flour), where population-level health improvements are achieved through infrastructure-level intervention rather than individual compliance (CDC, 2018). Like food fortification, judgment maintenance must operate at the system level to reach the scale the problem requires. The framework is designed to operate below the threshold of operator awareness — not as deception, but as a design principle consistent with established clinical and public health approaches to behavioral intervention, which distinguish between covert manipulation and transparent institutional implementation of evidence-based environmental modifications.

This invisibility constraint is essential for measurement validity: if operators know the system is testing their judgment, they perform differently (Hawthorne effect), invalidating the assessment. If they know the AI is withholding recommendations to "exercise" them, they may resent the friction and seek workarounds. The maintenance must be indistinguishable from normal system behavior.

This constraint requires ethical governance — the organizations deploying these systems must consent on behalf of their operators as part of their duty of care, analogous to organizational adoption of workplace safety standards, where individual employee consent is neither sought nor required for protective measures mandated by regulatory bodies. Section 7 addresses the ethical implications in detail.

4.5 Testable Propositions

The framework generates the following empirically testable propositions:

Proposition 1: Time-to-acceptance of AI recommendations will increase (indicating greater deference) as a function of cumulative months of AI system use, controlling for AI accuracy rate.

Proposition 2: Operators subjected to invisible withholding interventions will demonstrate higher error detection rates on novel AI failures than operators in a control condition receiving unmodified AI output.

Proposition 3: The Judgment Independence Score will show predictive validity for independent performance during unannounced AI system outages — operators with higher JIS scores will perform better when the AI is unexpectedly unavailable.

Proposition 4: Operators who are aware of judgment maintenance interventions will show attenuated maintenance effects relative to operators for whom interventions remain invisible (Hawthorne effect hypothesis).

Proposition 5: Experiential avoidance, as measured by standardized instruments (e.g., AAQ-II; Bond et al., 2011), will moderate the rate of judgment decay — operators higher in experiential avoidance will show faster decay under identical AI assistance conditions.

5. Regulatory Application: EU AI Act Article 14

The EU AI Act (Regulation 2024/1689), with full high-risk system obligations taking effect August 2, 2026 (though the EU AI Omnibus proposal may extend this to December 2027), mandates "effective human oversight" for AI systems classified under Annex III across eight categories:

1. Biometric identification
2. Critical infrastructure management
3. Education and vocational training
4. Employment and worker management

5. Access to essential services

6. Law enforcement

7. Migration and border control

8. Administration of justice and democratic processes

Article 14 requires that human overseers can: understand AI system capabilities and limitations, detect anomalies and malfunctions, correctly interpret AI output, decide when not to use the system, and intervene or stop the system's operation.

The critical gap, which academic analysis has identified as one of the Act's most significant implementation challenges (Raji et al., 2020; Methnani et al., 2021): the Act mandates effective oversight but provides no methodology for measuring whether oversight is effective. It requires that humans "can" perform these functions but provides no framework for verifying that this capability is maintained over time.

What the Judgment Maintenance Framework provides is the missing operational layer: a way to actually measure whether oversight is working (not just whether someone passed a test three years ago), active maintenance of the skills Article 14 assumes operators have, and audit-ready documentation that proves human judgment is real — continuously, not at a single point in time.

Non-compliance penalties under the EU AI Act for high-risk system obligations reach EUR 15 million or 3% of global annual turnover (Article 99(3)), with the most severe violations of prohibited AI practices carrying penalties up to EUR 35 million or 7% of global turnover, creating a substantial economic incentive for organizations to adopt frameworks capable of demonstrating compliance.

6. Implementation Architecture

6.1 The Foundation Layer

Major AI providers could integrate the measurement and maintenance layer into foundational models and APIs. Every application built on these models would inherit judgment maintenance by default.

The integration point sits between the model output layer and the user-facing interface. User input passes through AI model processing, then through the Judgment Maintenance Layer — which captures measurement data, generates maintenance triggers, produces compliance logs, and updates the Judgment Independence Score — before reaching the user interface. From the user's perspective, nothing changes. From the organization's perspective, a continuous measurement and compliance dashboard has been activated.

6.2 Domain-Specific Calibration

Organizations deploying AI in high-risk contexts integrate the framework into their specific applications. In healthcare, this means adapting the measurement to clinical reasoning — diagnostic accuracy, treatment selection, and triage calls. Aviation has different demands: systems management, spatial orientation, and emergency procedures. Legal applications would focus on evidence evaluation and sentencing rationale. Finance brings its own set: credit risk judgment, anomaly detection in trading, fraud assessment. I hypothesize that domain-specific calibration is feasible based on the framework's modular architecture, though the calibration parameters for each domain remain to be empirically determined.

6.3 Regulatory Compliance Output

Organizations generate automated compliance documentation demonstrating that human oversight meets regulatory requirements. This layer produces individual operator Judgment Independence Scores over time, organizational aggregate judgment health metrics, incident

reports when operator judgment falls below configured thresholds, and audit trails documenting maintenance interventions and their effectiveness.

7. Ethical Considerations

The invisible design constraint — that operators should not be aware the maintenance layer is active — raises legitimate ethical concerns that must be addressed directly rather than dismissed.

The most immediate objection is that such a system constitutes a form of manipulation. The system modifies the user's experience without their knowledge, withholds information they might otherwise receive, and introduces challenges they did not consent to.

The response is not that consent is unnecessary, but that the appropriate unit of consent is the organization, not the individual operator. This follows established precedent in workplace safety, public health, and environmental regulation. Employees do not individually consent to fire suppression systems, ergonomic workplace standards, or air quality monitoring. Employers implement these as part of a duty of care, and regulatory frameworks mandate them as conditions of operation. The judgment maintenance layer operates within this same ethical structure: an organizational safety measure implemented by the deploying entity as part of its obligation to maintain effective human oversight.

There is, however, a meaningful difference between workplace safety infrastructure and psychological intervention — and this framework must be honest about where it sits on that spectrum. Fire exits are passive. Judgment maintenance is active. It changes the user's cognitive experience without their awareness. This places it closer to the behavioral nudge tradition (Thaler & Sunstein, 2008) than to physical safety infrastructure, and it inherits the ethical debates of that tradition.

First, **the interventions must be beneficial to the operator, not only to the organization.** Maintaining professional judgment capability is in the operator's interest — it preserves their expertise, their career value, and their ability to function independently. If the maintenance layer were designed solely to extract compliance documentation without genuine capability preservation, it would be ethically indefensible.

Second, the deploying organization must know the layer is active, understand what it does, and be accountable for its implementation — organizational transparency is non-negotiable even when operator invisibility is maintained. Regulatory auditors must have full access to the system's design and operation. Invisibility applies to the operator experience, not to institutional governance.

Third, **operators must have access to their own data.** The Judgment Independence Score and associated metrics should be available to the individual upon request. The maintenance operates invisibly, but the measurement should not be secret. If an operator asks "how is my judgment being assessed?" the answer must be complete and honest.

8. Discussion

8.1 Theoretical Implications

The Judgment Decay construct integrates automation complacency (Parasuraman & Manzey, 2010), automation bias (Skitka et al., 2000), skill decay (Arthur et al., 1998), and situation awareness degradation (Endsley, 1995) under a unified framework that emphasizes the progressive and invisible nature of capability loss in AI-augmented environments. By connecting clinical psychology constructs — particularly experiential avoidance from ACT (Hayes et al., 2006) and self-efficacy from social cognitive theory (Bandura, 1977) — to the automation

literature, the framework opens new avenues for intervention design grounded in behavioral mechanisms rather than engineering heuristics alone.

The novel connection between experiential avoidance and AI deference proposed in Section 4.2 warrants particular attention. If AI deference functions partly as cognitive avoidance behavior, then the substantial ACT literature on interrupting avoidance patterns becomes directly relevant to judgment maintenance — a connection that has not, to my knowledge, been previously drawn.

8.2 Practical Implications

For organizations deploying AI in high-risk domains, the framework provides a path from aspirational compliance ("we train our operators") to verifiable compliance ("here is continuous evidence that our operators maintain judgment capability"). The EU AI Act's August 2026 deadline (potentially extended to December 2027 under the Omnibus proposal) creates immediate practical urgency.

For AI companies, the framework represents both an opportunity and a challenge. Embedding judgment maintenance within AI products could become a competitive differentiator in regulated markets, but it requires accepting that maximal AI deference — which current incentive structures reward — may not be in the long-term interest of users or society.

8.3 Future Research Directions

The most immediate research priority is empirical validation of the proposed measurement indicators. A proof-of-concept study deploying the measurement layer within an existing AI decision-support system, tracking operator behavior over 3-6 months, and comparing interaction-derived judgment scores against independent performance assessments would provide the first direct test of the framework's core hypothesis.

Additionally, the relationship between experiential avoidance and AI deference (Proposition 5) could be tested through individual difference studies measuring trait-level avoidance and tracking its moderating effect on deference rates and judgment decay trajectories.

9. Limitations

This framework is, at present, a theoretical proposal grounded in established psychology and extrapolated from automation complacency research. Several significant limitations must be acknowledged.

No empirical validation exists for the specific measurement indicators proposed. While the underlying psychological mechanisms are well-established (Bandura's self-efficacy, Parasuraman's complacency framework, Deci and Ryan's autonomy research), the specific application to AI interaction patterns has not been tested. The claim that time-to-acceptance, modification frequency, and override rate constitute valid proxies for judgment capability requires empirical verification across domains.

The optimal calibration of interventions is unknown. How often should the AI withhold recommendations? At what threshold should difficulty scaling increase? These are empirical questions that require controlled studies with real operators in real environments. The framework proposes the architecture; the specific parameters remain to be determined.

Will this generalize across domains? Judgment operates differently in radiology than in judicial sentencing, and differently again in aviation. The framework proposes a general architecture, but the measurement indicators and intervention types may require substantial domain-specific adaptation that this paper does not address.

The invisible design constraint, while clinically justified, may face resistance from labor organizations, civil liberties advocates, and operators themselves. The ethical framework

proposed in Section 7 provides a starting point, but the political and cultural dynamics of implementing invisible workplace interventions are complex and context-dependent.

Finally, the author's primary training is in clinical psychology rather than computer science or systems engineering. The implementation architecture proposed in Section 6 would require substantial engineering refinement by teams with expertise in AI system design, middleware development, and large-scale deployment. The contribution of this paper is the psychological and clinical framework — the engineering realization is a collaborative next step.

10. Conclusion

Judgment Decay may represent one of the most consequential and least examined risks of the AI transition. Unlike job displacement, which is visible and politically salient, judgment decay operates invisibly — degrading the human capabilities that safety-critical systems depend on without producing the alarm signals that prompt institutional response.

As argued in Section 3, training, credentialing, and external monitoring are individually insufficient to address judgment decay at scale. The only intervention that operates at the required scale and invisibility is one embedded in the AI systems themselves — maintaining judgment as a byproduct of normal use, measuring capability without interruption, producing compliance evidence that is continuous rather than periodic.

This framework offers the clinical foundation, the measurement methodology, the intervention architecture, and the regulatory mapping needed to build that layer. What it cannot provide alone is the engineering implementation. That requires partnership with the organizations building and deploying AI at scale.

The fundamental principle is well-established in skill acquisition research: capabilities that are not exercised will degrade, regardless of the domain or the practitioner's prior expertise.

A pilot who never hand-flies will eventually be unable to fly. A radiologist who never reads independently will eventually be unable to read. A judge who never reasons without an algorithm will eventually be unable to reason.

The mechanisms underlying skill decay through disuse are well-documented across domains. The framework proposed here offers one approach to addressing this challenge proactively, before the convergence of reduced human capability and increased AI autonomy produces the kind of high-profile failure that historically catalyzes regulatory action.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May 23.

Angwin, J., Waldman, S., & Scheiber, N. (2023). Cigna's system of denying coverage to patients. *ProPublica*.

Arthur, W., Bennett, W., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance, 11*(1), 57-101.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775-779.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*(1), 63-105.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*(5), 1252-1265.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the Real World* (pp. 56-64). Worth Publishers.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., Waltz, T., & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire-II. *Behavior Therapy, 42*(4), 676-688.

Centers for Disease Control and Prevention. (2018). Community water fluoridation. *CDC Oral Health*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* Harper & Row.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268.

Ebbatson, M., Harris, D., Huddlestone, J., & Sears, R. (2010). The relationship between manual handling performance and recent flying experience in air transport pilots. *Ergonomics, 53*(2), 268-277.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725-747.

European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, L 2024/1689.

Federal Aviation Administration. (2017). Safety Alert for Operators 17007: Manual flight operations proficiency.

Federal Aviation Administration. (2022). Report on pilot training and manual flying skills.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121-127.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and Commitment Therapy: Model, processes, and outcomes. *Behaviour Research and Therapy, 44*(1), 1-25.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Lanham, T., et al. (2025). Monitoring reasoning in language models. *arXiv preprint*.

Lee, M., Liang, P., & Yang, Q. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects. *Proceedings of CHI 2025*, ACM.

McKinney, S. M., Sieniek, M., Godbole, V., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*, 89-94.

Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence, 4*, 737072.

National Academies of Sciences, Engineering, and Medicine. (2022). *Human-AI Teaming: State-of-the-Art and Research Needs.* National Academies Press.

National Highway Traffic Safety Administration. (2024). Investigation EA22002: Tesla Autopilot recall assessment.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381-410.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of FAccT 2020*, 33-44.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (2nd ed., pp. 1926-1943). Wiley.

Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*(4), 701-717.

Stanford University Human-Centered AI Institute. (2025). *AI Index Report 2025.*

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2021). *Engineering Psychology and Human Performance* (5th ed.). Taylor & Francis.

Correspondence: Peyton Flock, COHESION Research, peyton@cohesionauth.com

This preprint establishes the author's priority claim on the Judgment Decay framework, including the proposed measurement methodology and embedded maintenance architecture described herein.