# Human–AI Handovers: A Dynamic Authority Reversal Framework for Trust Calibration and Transitional Accountability

Jonathan H. Westover *

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0390.v1

Keywords: human–AI collaboration; authority transition; transitional accountability; trust calibration; explainable AI; algorithmic governance; adaptive automation

 doi:10.20944/preprints202603.0390.v1



*Article*

# Human–AI Handovers: A Dynamic Authority Reversal Framework for Trust Calibration and Transitional Accountability

**Jonathan H. Westover**

Nexus Institute for Work & AI – Catalyst Center for Work Innovation; jon.westover@gmail.com

**Abstract**

Human–artificial intelligence collaboration is increasingly treated as a static allocation problem—humans decide, machines compute—yet high-stakes workflows reveal a more fluid reality: leadership shifts multiple times within a single decision episode. This paper formalizes the **Dynamic Authority Reversal (DAR)** framework, which models intra-episode authority transitions across four states: Human-Leader/AI-Follower (HL), AI-Leader/Human-Follower (AL), Co-Leadership (CO), and Mutual Override (MO). Transitions are governed by four trigger classes—data superiority, contextual judgment requirements, risk thresholds, and ethics overrides—and are stabilized through hysteresis bands and safe-exit timers. The framework couples micro-level trust calibration with macro-level legitimacy by introducing the Reversal Register, an auditable log that binds each decision to the prevailing authority state, trigger conditions, and justificatory explanations. Ten falsifiable propositions are derived and linked to measurement constructs, prioritized by foundational importance and empirical tractability. Sector-specific implementation guidance is provided for healthcare and public administration, with attention to existing governance structures and regulatory frameworks. By operationalizing handovers rather than merely prescribing "human oversight," DAR advances both theory and practice: it equips researchers with testable hypotheses, furnishes practitioners with governance-ready instruments, and offers regulators an auditable architecture that preserves ultimate human accountability while enabling reversible AI leadership where contextually advantageous.

**Keywords:** human–AI collaboration; authority transition; transitional accountability; trust calibration; explainable AI; algorithmic governance; adaptive automation

---

## 1. Introduction

### *1.1. The Therac-25 Catastrophe as Parable*

Between 1985 and 1987, a computerized radiation therapy machine called the Therac-25 delivered massive overdoses to at least six patients, causing deaths and severe injuries. Unlike earlier models, the Therac-25 removed hardware safety interlocks and relied entirely on software to prevent dangerous beam configurations. When a specific sequence of rapid operator keystrokes occurred, the software failed to detect that the machine was in the wrong mode, delivering radiation doses hundreds of times greater than intended (Leveson & Turner, 1993). The operators—highly trained technicians—trusted the machine's displayed readings and had no mechanism to recognize that the system had entered an unsafe state. By the time the pattern was identified, lives had been lost.

The Therac-25 catastrophe is instructive not because automation failed in isolation, but because the *handover* between machine autonomy and human oversight was fundamentally misconceived. The system's designers assumed that software could safely replace hardware interlocks, yet they provided operators with no meaningful signals that would indicate when human intervention was

 doi:10.20944/preprints202603.0390.v1



required. Operators, in turn, had been trained to trust machine readings implicitly, with no protocol for questioning or overriding them.

A clarification is warranted: the Therac-25 failure was primarily an *inter-system* design failure—the removal of hardware interlocks in favor of software-only safeguards. However, the tragedy also illuminates *intra-episode* dynamics: operators cycling through treatment setups had no mechanism to detect when their rapid inputs had placed the system in an unsafe state, no signal prompting them to reassert authority, and no protocol for intervention. The catastrophe thus crystallizes a truth that remains urgent today: in high-stakes human–machine collaboration, the question is not merely *whether* humans retain oversight but *how* authority transitions are designed, signaled, and governed across the temporal unfolding of a decision episode.

### *1.2. The Limits of Static Role Taxonomies*

Prevailing conceptualizations of human–AI interaction rest on a taxonomy of static roles. *Augmentation* positions AI as a tool that extends human capacity without displacing human judgment (Raisch & Krakowski, 2021). *Automation* transfers discrete tasks to machines under the assumption that human oversight can be invoked when needed. *Human-in-the-loop* (HITL) designs insert human checkpoints at critical junctures but often leave unspecified the signals that should trigger those checkpoints or the procedures that should govern them (Amershi et al., 2019). Each category presupposes that roles, once assigned, remain fixed for the duration of a task or workflow.

Yet field evidence contradicts this assumption. In algorithmic trading, leadership oscillates between quantitative models and human traders as volatility regimes shift. In radiology, AI may dominate routine screening while human radiologists assume primacy when images present ambiguous pathology (Topol, 2019). In child-welfare screening, predictive-risk tools may inform—but not determine—caseworker judgment, with the locus of authority negotiated case by case (Eubanks, 2018). These examples share a common structure: authority is not partitioned once but *exchanged repeatedly* within a decision episode, often in response to real-time signals about data quality, contextual complexity, risk magnitude, or normative stakes.

### *1.3. The Need for Transition Logic*

If authority is dynamic, organizations require a *transition logic*—a principled account of who leads at each moment, why leadership changes, and how accountability attaches to the party in charge. Without such logic, three failure modes become likely:

1. **Over-control.** Organizations cling to human primacy even when AI demonstrably outperforms, sacrificing accuracy and efficiency.
2. **Under-control.** Organizations cede authority to AI without specifying conditions for reversion, incurring safety, legitimacy, and compliance risks.
3. **Accountability voids.** Decisions fall into gaps between human and machine responsibility, leaving neither party answerable when outcomes go awry.

The Therac-25 tragedy exemplified the second and third failure modes simultaneously: authority had been ceded to software without meaningful reversion mechanisms, and when catastrophe struck, accountability was diffused across hardware engineers, software developers, hospital administrators, and regulatory bodies—none of whom had clear responsibility for the handover failures that enabled the overdoses.

Emerging regulatory regimes underscore the ongoing stakes. The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) mandates documented human oversight for high-risk AI applications, yet the statute's language remains at the level of principle rather than mechanism. Regulators demand that humans retain "meaningful control," but what counts as meaningful—and how control is operationalized during intra-episode handovers—remains under-specified.

### *1.4. Introducing the Dynamic Authority Reversal Framework*

 doi:10.20944/preprints202603.0390.v1



This paper addresses the foregoing lacunae by formalizing the **Dynamic Authority Reversal (DAR)** framework. The term "reversal" is chosen deliberately: while not all transitions represent returns to prior states, the framework emphasizes the *reversibility* of authority—the principled capacity to reclaim or reassign leadership as conditions evolve. DAR reconceptualizes human–AI collaboration as a *temporal authority problem* in which leadership shifts across four states:

- **Human-Leader/AI-Follower (HL):** Humans decide; AI advises.
- **AI-Leader/Human-Follower (AL):** AI proposes and, absent override, enacts; humans monitor.
- **Co-Leadership (CO):** Authority is negotiated or shared via explicit merge protocols; neither party holds unilateral control.
- **Mutual Override (MO):** A protective interrupt—either party may halt or reverse the process when risk or ethical thresholds are breached.

Transitions between states are triggered by four classes of signals: *data superiority* (when AI predictive performance exceeds a defined margin), *contextual judgment requirements* (when tacit knowledge or relational factors predominate), *risk thresholds* (when potential harms exceed tolerance), and *ethics overrides* (when normative constraints—legal, professional, or organizational—demand human primacy). *Hysteresis bands* prevent oscillation by requiring that exit thresholds differ from entry thresholds. *Safe-exit timers* guarantee human review after a maximum dwell time in any state (Regulation (EU) 2024/1689). The *Reversal Register* logs state, trigger, actors, thresholds, explanations delivered, and actions taken, creating an auditable trail that links each decision to its prevailing authority configuration.

### *1.5. Contributions*

DAR advances theory and practice along four dimensions:

1. **Transition logic.** DAR specifies *when* and *why* authority should shift, moving beyond the static categories that dominate existing literature and integrating insights from the adaptive-automation tradition.
2. **Transitional accountability.** DAR couples micro-level trust calibration with macro-level legitimacy, theorizing accountability along multiple dimensions (answerability, responsibility, liability) and binding each to authority states.
3. **Operational instrumentation.** DAR translates theoretical constructs into buildable artifacts: Authority-State Playbooks, safe-exit timers, state-contingent explanation templates, and telemetry specifications that make handovers auditable and measurable.
4. **Attention to human experience.** DAR acknowledges that operators experience handovers phenomenologically—as potentially disorienting, contested, or empowering—and incorporates design considerations to support the human experience of transitions.

The remainder of the paper proceeds as follows. Section 2 situates DAR within cognate literatures on human–AI interaction, adaptive automation, leadership, trust, explainability, and accountability. Section 3 presents the DAR model in formal detail, including theoretical justification for the state taxonomy and governance mechanisms for trigger disputes. Section 4 derives ten falsifiable propositions with associated measurement constructs, prioritization, and methodological pathways. Section 5 offers in-depth sector-specific implementation guidance for healthcare and public administration. Section 6 discusses theoretical implications, practical recommendations, political dimensions, the experience of affected publics, and boundary conditions. Section 7 concludes.

## 2. Literature Review

### *2.1. Human–AI Interaction: From Static Roles to Dynamic Handovers*

#### 2.1.1. The Augmentation–Automation Continuum

 doi:10.20944/preprints202603.0390.v1



Early theorizing on human–machine teaming adopted an augmentation–automation continuum. Augmentation treats technology as a cognitive prosthesis that amplifies human capability without supplanting human judgment (Davenport & Kirby, 2016). Automation transfers task execution to machines, reserving for humans a supervisory or exception-handling role. Both poles assume relatively stable role assignments: either the human decides with machine support, or the machine decides with human fallback.

Empirical research has refined this picture by introducing intermediate categories. Parasuraman and colleagues' influential taxonomy distinguishes levels of automation across four generic functions—information acquisition, information analysis, decision selection, and action implementation—and within each function ranks automation from low (human performs) to high (machine performs) (Parasuraman et al., 2000). The taxonomy has proven durable, yet it remains essentially structural: it classifies *what* is automated but not *when* or *how* automation levels should change during task performance.

### 2.1.2. Adaptive Automation and Dynamic Function Allocation

Recognizing the limitations of static allocation, researchers developed the concept of *adaptive automation*—systems that dynamically reallocate functions between humans and machines based on real-time assessment of operator state, task demands, or environmental conditions (Kaber & Endsley, 2004). Sheridan and Verplank's (1978) foundational taxonomy of supervisory control anticipated this development by distinguishing degrees of computer autonomy, from full human control to full computer control, with intermediate levels involving computer recommendation, execution pending human approval, or execution with human veto.

Subsequent work in robotics and autonomous systems has elaborated adaptive-autonomy frameworks. Goodrich and Schultz (2007) review sliding-autonomy approaches in human–robot interaction, where the locus of control shifts based on task phase, robot confidence, or operator workload. These frameworks share with DAR an emphasis on dynamic allocation but typically focus on human–robot dyads in physical environments rather than on the broader organizational and accountability contexts of AI-assisted decision-making.

DAR extends the adaptive-automation tradition in three ways. First, it generalizes from operator-state triggers (e.g., workload, fatigue) to a broader trigger taxonomy encompassing data superiority, contextual judgment, risk, and ethics. Second, it foregrounds accountability by introducing the Reversal Register and the concept of transitional accountability. Third, it integrates insights from organizational theory—particularly distributed leadership and trust calibration—that are absent from engineering-focused adaptive-automation models.

### 2.1.3. Human-in-the-Loop and Its Discontents

The HITL paradigm inserts human checkpoints at critical decision junctures, aiming to combine algorithmic efficiency with human judgment. Design guidelines for HITL systems emphasize legibility, controllability, and graceful degradation (Amershi et al., 2019). Yet field deployments reveal persistent difficulties. Operators may treat algorithmic recommendations as de facto decisions, exercising nominal rather than substantive oversight—a phenomenon labeled *automation complacency* (Parasuraman & Manzey, 2010). Conversely, operators may distrust accurate algorithms and override them inappropriately, incurring *algorithm aversion* (Dietvorst et al., 2015). Neither pathology is addressed by static role assignments; both reflect a failure to calibrate authority dynamically to context.

Research on mode confusion and automation surprises further underscores the challenges of human–automation interaction. Sarter and Woods (1995) documented how pilots interacting with automated cockpit systems frequently lost awareness of which mode the automation was in, leading to errors and near-misses. These findings highlight that even well-designed transitions can produce disorientation if operators lack adequate feedback about authority state.

 doi:10.20944/preprints202603.0390.v1



Recent work has begun to recognize the need for adaptive oversight. Research on algorithmic decision-making demonstrates that the benefits of algorithmic advice depend on how advice is framed and on whether operators can adjust their reliance over time (Green & Chen, 2019). Studies examining human predictions with AI support show that user trust in AI predictions evolves with experience but may over- or under-shoot appropriate levels (Lai & Tan, 2019). These findings imply that optimal human–AI collaboration is not a fixed configuration but a moving target—one that demands continuous recalibration of who leads and who follows.

#### 2.1.4. Toward Transition Logic

What is missing from the foregoing literature is a *transition logic* that integrates adaptive-automation mechanisms with organizational accountability structures: a specification of the signals that should prompt authority shifts, the procedures that should govern those shifts, the accountability structures that should attend them, and the governance processes that should oversee threshold-setting. DAR fills this gap by conceptualizing authority as a state variable and by articulating principled triggers, guardrails, telemetry, and governance mechanisms for state transitions.

### *2.2. Leadership Theory: Distributed, Shared, and Hybrid Forms*

#### 2.2.1. Distributed and Shared Leadership

Leadership scholarship has long recognized that influence need not concentrate in a single individual. *Distributed leadership* distributes influence functions across organizational members according to expertise and situational demands (Gronn, 2002). *Shared leadership* emphasizes mutual influence among team members who collectively steer toward common goals (Pearce & Conger, 2003). Meta-analytic evidence confirms that shared leadership is positively associated with team performance, satisfaction, and viability, particularly in knowledge-intensive and dynamic environments (D'Innocenzo et al., 2016).

These literatures offer conceptual resources for thinking about human–AI teams. If leadership can be distributed among humans, it can, in principle, be distributed between humans and machines—or so the analogy suggests. Recent work has extended shared-leadership constructs to human–AI collaboration, arguing that algorithms can occupy follower, partner, or even leader roles depending on task characteristics and team configurations (Seeber et al., 2020). However, shared-leadership research has not yet specified the mechanisms by which leadership rotates or the conditions under which rotation should occur. DAR contributes such mechanisms.

#### 2.2.2. Leadership as Influence Exchange

A complementary perspective treats leadership as an *influence exchange* in which parties claim and grant influence based on perceived competence, legitimacy, and situational fit (DeRue & Ashford, 2010). This relational view suggests that leadership is not a fixed attribute but an emergent property of interaction. Extending the framework to human–AI contexts, one might ask: Under what conditions does an AI system's demonstrated competence warrant granting it influence? When does contextual ambiguity or normative complexity warrant reclaiming influence for humans?

DAR operationalizes these questions by defining triggers—data superiority, contextual judgment requirements, risk thresholds, ethics overrides—that translate situational signals into authority transitions. In so doing, it bridges leadership theory's relational insights and the operational demands of sociotechnical system design.

### *2.3. Trust Calibration in Human–AI Interaction*

#### 2.3.1. The Trust-Calibration Challenge

 doi:10.20944/preprints202603.0390.v1



Trust is a psychological state in which an agent accepts vulnerability based on positive expectations of another's behavior (Mayer et al., 1995). In human–AI contexts, trust calibration refers to aligning an operator's reliance on an AI system with the system's actual reliability (Lee & See, 2004). Miscalibration manifests as *over-trust* (relying on AI when it errs) or *under-trust* (overriding AI when it is correct). Both pathologies compromise decision quality.

Empirical research documents the prevalence of miscalibration. Studies find that operators often anchor on initial impressions of AI reliability and update slowly in the face of contradictory evidence (Yin et al., 2019). Transparency interventions—providing explanations, confidence scores, or performance feedback—can improve calibration under some conditions but may backfire under others, for example by overwhelming operators with information or by inducing false confidence (Poursabzi-Sangdeh et al., 2021).

#### 2.3.2. Dynamic Trust and Authority

DAR reframes the trust-calibration problem by linking it to authority transitions. Rather than asking, *How much should the operator trust the AI?* DAR asks, *Who should lead given current conditions, and how should trust-relevant information be presented to support appropriate reliance?* This reframing has practical implications. When AI leads (AL state), explanations should emphasize operational sufficiency—confidence bounds, alternative options, risk envelopes—so that human monitors can detect when reversion is warranted. When humans lead (HL state), explanations should summarize AI recommendations concisely to inform but not displace human judgment. When authority is shared (CO state), explanations should highlight areas of convergence and divergence to facilitate joint deliberation. When override conditions obtain (MO state), explanations should justify the override to preserve accountability.

Research on explanation effects in AI-assisted decision-making provides supporting evidence, demonstrating that the way information is presented affects trust calibration and that calibration improves when explanations match operator needs at specific decision phases (Yin et al., 2019). DAR generalizes this insight by tying explanation strategy to authority state.

### *2.4. Explainable AI: From Transparency to State-Contingent Legibility*

#### 2.4.1. The Explainability Landscape

Explainable AI (XAI) has emerged as a response to the opacity of complex machine-learning models. Explanation methods range from feature-attribution techniques (e.g., SHAP, LIME) to counterfactual reasoning to natural-language rationales (Miller, 2019). A burgeoning literature evaluates explanation quality along dimensions such as fidelity, comprehensibility, and actionability (Liao et al., 2020).

Yet a recurrent finding is that *more* explanation does not automatically yield *better* human–AI performance. Bansal and colleagues (2021) show that explanations can improve complementary team performance—where the human–AI team outperforms either party alone—but only under specific conditions; poorly designed explanations may induce over-reliance or confusion. Buçinca and colleagues (2021) find that cognitive-forcing interventions, which require operators to commit to a judgment before seeing AI recommendations, can mitigate over-reliance more effectively than transparency alone.

#### 2.4.2. Explanations as Phase-Specific Instruments

DAR's contribution to the explainability discourse is to treat explanations as *phase-specific instruments* rather than generic documentation. The framework specifies explanation classes by authority state:

- **HL state:** Summative explanations that distill AI recommendations for human evaluation.
- **AL state:** Operational-sufficiency explanations that convey confidence, alternatives, and risk boundaries.

 doi:10.20944/preprints202603.0390.v1



- **CO state:** Diagnostic explanations that illuminate areas of agreement and disagreement.
- **MO state:** Override justifications that document the rationale for halting or reversing action.

This differentiation aligns with Liao and colleagues' (2020) call for user-centered explanation design and with Miller's (2019) insight that effective explanations are contrastive, selective, and social. By indexing explanation strategy to authority state, DAR makes explanation design a tractable engineering problem rather than a one-size-fits-all aspiration.

### *2.5. Accountability and Legitimacy in Algorithmic Governance*

#### 2.5.1. The Accountability Gap

A persistent concern in algorithmic governance is the diffusion of responsibility. When AI systems inform or effectuate consequential decisions—credit denials, medical diagnoses, benefit determinations—affected parties and regulators may struggle to identify who is answerable for outcomes. Raji and colleagues (2020) characterize this as an "accountability gap" and propose internal algorithmic-auditing frameworks to close it. Mittelstadt (2019) observes that high-level ethical principles (fairness, transparency, accountability) often lack operational mechanisms, leaving gaps between aspiration and implementation.

#### 2.5.2. Dimensions of Accountability

Accountability is a multidimensional construct. Bovens (2007) distinguishes several components: *answerability* (the obligation to explain and justify conduct), *responsibility* (the duty to act appropriately), and *liability* (the exposure to sanctions for failures). Wieringa (2020) extends this framework to algorithmic contexts, arguing that effective algorithmic accountability requires mechanisms that address all three dimensions.

DAR's concept of transitional accountability engages each dimension:

- **Answerability:** The Reversal Register documents which party held authority, what triggers prompted transitions, and what explanations were provided, enabling retrospective justification of decisions.
- **Responsibility:** By binding authority to states, DAR clarifies who bears the duty to act appropriately at each moment—the human in HL, the human monitor in AL, both parties under explicit protocols in CO.
- **Liability:** When outcomes go awry, the Reversal Register provides an evidentiary basis for attributing liability to the party in charge at the relevant moment, subject to whether that party discharged their duties appropriately.

This theorization clarifies that transitional accountability does not collapse accountability dimensions but rather operationalizes each within a dynamic authority framework.

#### 2.5.3. Transitional Accountability in Practice

Rather than assigning responsibility to "the algorithm" or "the operator" in the abstract, DAR binds responsibility to the authority state prevailing at the moment of decision. The Reversal Register documents, for each decision: (a) the current state, (b) the trigger that prompted entry into that state, (c) the thresholds in effect, (d) the explanations delivered, and (e) the actions taken by each party. This record enables auditors and regulators to trace responsibility with precision: if the AI led and the human monitor failed to override despite a detectable signal, the monitor bears responsibility for the lapse; if the system failed to surface a requisite signal, system designers bear responsibility for the omission.

The approach resonates with sociotechnical critiques that emphasize the importance of institutional context for algorithmic accountability (Selbst et al., 2019). DAR does not abstract away context; rather, it operationalizes context through state definitions, trigger conditions, and logging protocols. By making authority transitions visible and auditable, DAR links micro-level operator

 doi:10.20944/preprints202603.0390.v1



behavior to macro-level governance structures, satisfying regulatory expectations for meaningful human oversight (Regulation (EU) 2024/1689).

### *2.6. Synthesis and Research Gap*

The foregoing literatures converge on a recognition that human–AI collaboration is dynamic, that static role taxonomies are insufficient, and that accountability requires traceable mechanisms. The adaptive-automation tradition has developed dynamic allocation mechanisms but has not addressed organizational accountability contexts. Leadership theories acknowledge distributed influence but not the triggers that should prompt reallocation. Trust research documents calibration challenges but not the structural interventions that tie calibration to authority assignment. XAI research develops explanation methods but not the principle that explanation strategy should vary with authority state. Accountability theory distinguishes dimensions of accountability but lacks operational mechanisms for dynamic human–AI contexts.

DAR fills this gap by:

1. Specifying four authority states with clear semantics and diagnostic criteria.
2. Defining four trigger classes, governance mechanisms for trigger disputes, and attendant guardrails.
3. Introducing instrumentation—safe-exit timers, hysteresis bands, Reversal Registers—that operationalizes transitions.
4. Theorizing transitional accountability along multiple dimensions.
5. Deriving falsifiable propositions that enable cumulative research.

## 3. The Dynamic Authority Reversal Model

### *3.1. Overview*

The DAR model conceptualizes human–AI collaboration as a finite-state system in which leadership—defined as the party with primary authority to propose and, absent objection, enact decisions—varies over the course of a decision episode. The model comprises four states, four trigger classes, temporal guardrails, governance mechanisms, and a logging mechanism.

### *3.2. Theoretical Justification for the State Taxonomy*

#### 3.2.1. Deriving States from First Principles

The four-state taxonomy is derived from the intersection of two dimensions: (a) *locus of initiative* (who proposes action) and (b) *locus of enactment authority* (who can effectuate action absent objection). Crossing these dimensions yields four configurations:

| | **Human Enacts** | **AI Enacts (absent override)** |
|---|---|---|
| **Human Proposes** | HL (Human-Leader) | — |
| **AI Proposes** | CO (negotiated merge) | AL (AI-Leader) |
| **Either Proposes Halt** | MO (protective interrupt) | MO (protective interrupt) |

- **HL:** Human proposes and enacts; AI advises.
- **AL:** AI proposes and enacts (absent override); human monitors.
- **CO:** AI proposes, but enactment requires human approval or negotiated merge—neither party holds unilateral authority.
- **MO:** Either party proposes halt; enactment is suspended pending resolution.

This derivation clarifies that the taxonomy is not arbitrary but reflects the logically distinct configurations of initiative and enactment authority in a two-party system with reversibility.

#### 3.2.2. MO as Meta-Level Interrupt

 doi:10.20944/preprints202603.0390.v1



Reviewer 1 raised the question of whether MO should be modeled as a coordinate state or as a meta-level interrupt. The derivation above suggests a middle position: MO is a *state* in the sense that the system occupies it during the halt, but it is *triggered* from any other state when override conditions obtain. Functionally, MO operates as a protective interrupt that suspends normal operation; semantically, it is the authority configuration during the suspension. This dual character is captured by representing MO as a state with incoming arcs from all other states and outgoing arcs to HL (the default safe state) or, following resolution, to any other state.

### *3.3. States and Diagnostic Criteria*

To address concerns about boundary ambiguity, this section provides operational definitions and diagnostic criteria for each state.

#### 3.3.1. Human-Leader/AI-Follower (HL)

*Definition:* The human operator holds decision authority. AI contributes recommendations, analyses, or confirmatory signals, but the human retains the final say and enacts the decision.

*Diagnostic criteria:*

- The human must actively approve, modify, or reject the AI recommendation before action proceeds.
- The AI cannot enact action unilaterally.
- The human's approval is not merely nominal (e.g., a one-click confirmation after AI has already acted) but substantive (e.g., review of recommendation with opportunity to modify).

*Accountability locus:* Human operator.

*Explanation target:* Summative presentation of AI recommendations sufficient for informed human judgment.

*Example (healthcare):* A physician reviews an AI-generated differential diagnosis, considers patient history and clinical intuition, and determines the treatment plan. The AI's recommendation is input; the physician's decision is output.

#### 3.3.2. AI-Leader/Human-Follower (AL)

*Definition:* The AI system proposes a course of action and, if not overridden within a defined window, enacts it. A designated human monitor observes and retains veto authority.

*Diagnostic criteria:*

- The AI can enact action if the human does not intervene within the designated window.
- The human's role is monitoring and exception-handling, not routine approval.
- The human has access to sufficient information (explanations, alerts) to detect when override is warranted.

*Accountability locus:* The designated human owner remains ultimately responsible; operational lead is AI.

*Explanation target:* Operational sufficiency—confidence bounds, alternative options, risk envelope—so that the monitor can judge whether override is warranted.

*Example (finance):* An algorithmic trading system executes orders autonomously within pre-set risk limits; a human risk officer monitors positions and can halt trading if limits are approached.

#### 3.3.3. Co-Leadership (CO)

*Definition:* Authority is negotiated or shared via an explicit merge protocol. Neither party holds unilateral control; decisions emerge from joint deliberation or from a defined procedure that combines human and AI inputs.

*Diagnostic criteria:*

- An explicit merge protocol is documented (e.g., weighted voting, sequential deliberation, consensus requirement).

 doi:10.20944/preprints202603.0390.v1



- Neither party can enact action without the other's input or approval.
- Disagreements are resolved through the documented protocol, not unilateral override.

*Distinguishing CO from AL:* In AL, the human can override but the default is AI enactment; in CO, there is no default—enactment requires joint action under the merge protocol.

*Accountability locus:* Shared, with explicit division of labor documented in the merge protocol.

*Explanation target:* Diagnostic presentation of areas of agreement and disagreement to facilitate synthesis.

*Example (public administration):* A social-welfare caseworker and a predictive-risk tool jointly assess a family's support needs using a structured protocol: the caseworker reviews the tool's score, adds observational evidence, and the two inputs are combined via a documented weighting scheme to determine the service plan.

#### 3.3.4. Mutual Override (MO)

*Definition:* A protective interrupt state activated when either party detects that risk or ethical thresholds are breached. Normal operation is suspended pending resolution.

*Diagnostic criteria:*

- An override condition has been signaled by either party.
- Action is halted; no enactment occurs until the condition is resolved.
- Resolution proceeds via a documented escalation protocol.

*Accountability locus:* The party invoking override assumes responsibility for the halt; ultimate accountability remains with the human.

*Explanation target:* Override justification—articulation of the condition that triggered MO and the rationale for halting.

*Example (autonomous vehicles):* A human safety driver disengages autonomous mode when a sensor anomaly is detected; alternatively, the vehicle's system initiates a safe stop when operating-domain boundaries are exceeded.

#### 3.3.5. Worked Example: Tracing a Decision Episode

Consider a radiology workflow in which an AI system screens chest X-rays for pneumonia (see Figure 1):

1. **Initial state (HL):** The radiologist reviews the AI's flagged cases, examining images and AI confidence scores. For each case, the radiologist makes the diagnostic determination. *Diagnostic criteria satisfied:* Human must approve before diagnosis is recorded.
2. **Transition to AL (data-superiority trigger):** Retrospective audit shows that for routine, high-confidence cases (AI confidence > 0.95), AI accuracy exceeds radiologist accuracy by 8% over 500 cases. Per the Authority-State Playbook, the workflow transitions to AL for cases meeting these criteria. *Diagnostic criteria satisfied:* AI diagnosis is recorded unless the radiologist overrides within 60 seconds.
3. **Transition to MO (risk-threshold trigger):** The AI flags a case with confidence 0.97, but the patient's electronic health record indicates a history of atypical presentations. The system's out-of-distribution detector fires, triggering MO. *Diagnostic criteria satisfied:* Diagnosis is suspended; escalation protocol initiates.
4. **Resolution to HL:** The radiologist reviews the case, determines that the atypical history warrants full human evaluation, and the system reverts to HL for this case.

This example illustrates how states are distinguished in practice and how triggers govern transitions.

 doi:10.20944/preprints202603.0390.v1



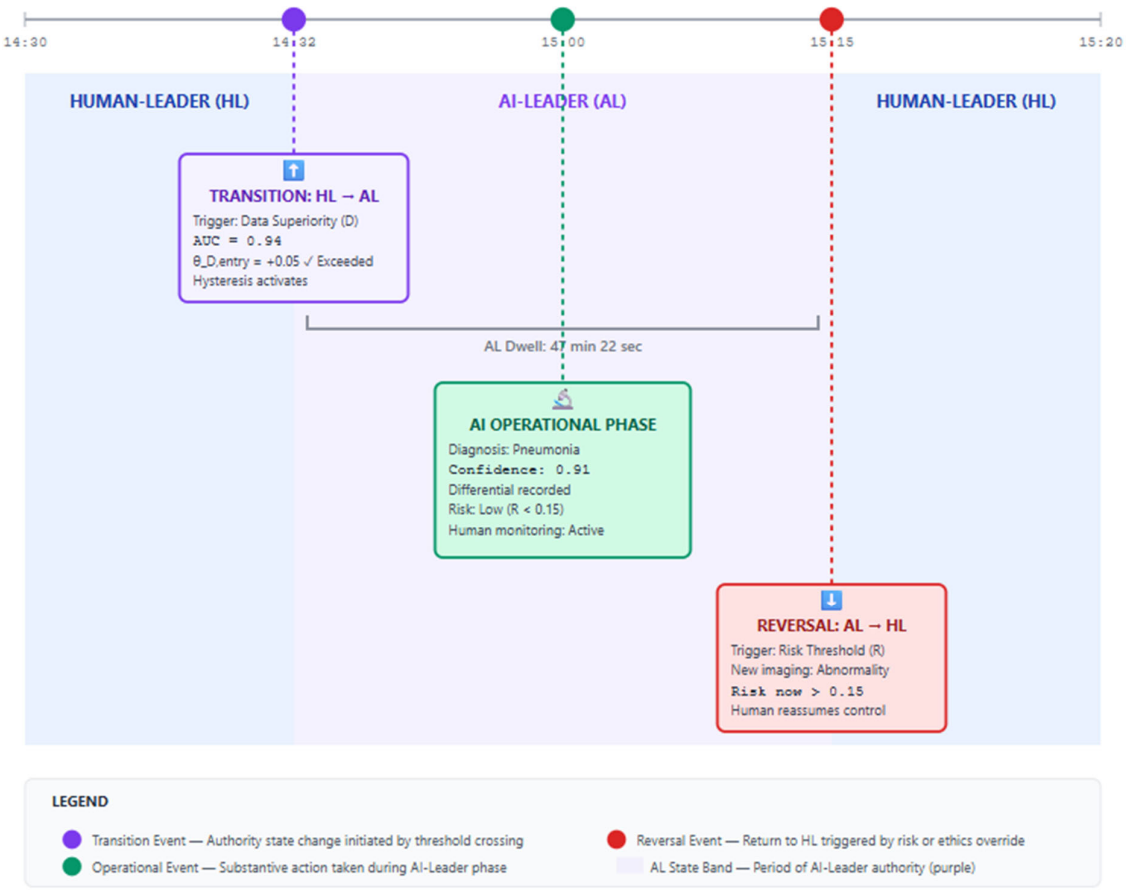


**Figure 1.** Worked Example Timeline.

### *3.4. Trigger Classes*

Transitions between states are governed by four classes of signals.


**Figure 2.** Trigger–Transition Matrix.

#### 3.4.1. Data Superiority (D)

A transition toward greater AI authority (e.g., HL → AL or HL → CO) may be warranted when AI predictive performance exceeds a defined margin over a baseline (e.g., human-only or prior-model

 doi:10.20944/preprints202603.0390.v1



performance) for a sustained period. The margin and duration are design parameters, calibrated to domain-specific performance metrics.

*Operationalization:* Define a performance metric $P$ (e.g., AUC, F1, profit factor). Let $\theta\_D$ denote the superiority threshold (e.g., $P\_AI > P\_baseline + 0.05$) and $\tau\_D$ the duration threshold (e.g., 100 consecutive decisions or 24 hours). When both conditions are satisfied, the trigger fires.

*Challenges and governance:* Baselines may shift, and performance metrics are often contested. Organizations should establish a *Trigger Governance Committee* (see Section 3.6) to define metrics, adjudicate disputes, and conduct periodic recalibration.

#### 3.4.2. Contextual Judgment Requirements (C)

A transition toward greater human authority (e.g., AL → HL or CO → HL) may be warranted when the decision context involves tacit knowledge, relational considerations, or novel situations not represented in training data.

*Operationalization:* Define an uncertainty metric $U$ (e.g., predictive entropy, distance from training-data manifold). Let $\theta\_C$ denote the threshold above which contextual judgment is required. When $U > \theta\_C$, the trigger fires.

*Challenges and governance:* High model uncertainty does not always indicate human superiority; it may indicate hard cases where both struggle. Organizations should validate that C-trigger cases empirically benefit from human authority and adjust thresholds accordingly.

#### 3.4.3. Risk Threshold (R)

A transition to MO—or reversion to HL—may be warranted when potential harms exceed a defined tolerance.

*Operationalization:* Define a risk metric $R$ and a threshold $\theta\_R$. When $R > \theta\_R$, the trigger fires. Thresholds may be absolute (e.g., loss exceeding $X, mortality probability exceeding Y%) or relative (e.g., risk index exceeding the 99th percentile of historical distribution).

#### 3.4.4. Ethics Override (E)

A transition to MO—or reversion to HL—may be mandatory when normative constraints are implicated: statutory requirements, professional codes, organizational values, or fundamental-rights considerations.

*Operationalization:* Define a set of normative constraints $E$ = {e_1, e_2, …, e_n}. If any $e\_i$ is implicated, the trigger fires. Implementation may involve rule-based flags, human annotation, or classification models trained on regulatory corpora.

*Challenges and governance:* Defining $E$ presupposes organizational consensus that may not exist. The Trigger Governance Committee should include representatives from legal, compliance, ethics, and affected stakeholder groups to negotiate the constraint set and adjudicate disputes.

### *3.5. Guardrails*

To prevent instability and to preserve meaningful human oversight, DAR incorporates two guardrails.

#### 3.5.1. Hysteresis

*Hysteresis* imposes asymmetric thresholds for entering and exiting a state. For example, the threshold for transitioning from HL to AL (data-superiority entry) may be $\theta\_D,entry = +0.07$, while the threshold for reverting from AL to HL (data-superiority exit) may be $\theta\_D,exit = +0.03$. The gap prevents oscillation ("thrashing") when performance hovers near a boundary.

*Design parameter:* Hysteresis width $\Delta\_D = \theta\_D,entry - \theta\_D,exit$. Wider hysteresis reduces thrash rate but may delay warranted reversals; narrower hysteresis is more responsive but risks instability.

 doi:10.20944/preprints202603.0390.v1



*Parameter-setting guidance:* Initial hysteresis widths should be set based on historical performance volatility. Simulation studies (see Proposition 3) can identify optimal widths for a given domain. Organizations should monitor thrash rates and reversal latencies and adjust widths during continuous-improvement cycles.

#### 3.5.2. Safe-Exit Timers

*Safe-exit timers* guarantee that no state persists indefinitely without human review. For each state, a maximum dwell time is defined; upon expiration, the system triggers a mandatory human checkpoint. Safe-exit timers operationalize the regulatory expectation that human oversight be *meaningful* rather than nominal (Regulation (EU) 2024/1689).

*Design parameter:* Timer duration $t\_max$. Shorter timers impose more frequent checkpoints (higher assurance, higher operator load); longer timers reduce load but risk prolonged autonomy without review.

*Parameter-setting guidance:* Timer durations should reflect workflow cadence (e.g., end-of-shift review, end-of-day review) and risk tolerance. High-stakes workflows warrant shorter timers; routine workflows may tolerate longer timers.

### *3.6. Trigger Governance*

Because threshold-setting is inherently political—lower data-superiority thresholds empower AI; higher thresholds empower humans—organizations should establish a *Trigger Governance Committee* with the following responsibilities:

1. **Threshold calibration.** Define initial thresholds based on pilot data and domain expertise.
2. **Dispute resolution.** Adjudicate disagreements about whether a trigger has fired appropriately.
3. **Periodic review.** Recalibrate thresholds at defined intervals (e.g., quarterly) based on performance data and stakeholder input.
4. **Stakeholder representation.** Include representatives from operations, legal, compliance, ethics, and—where feasible—affected publics.

The Committee's decisions should be documented and subject to internal audit, ensuring that threshold governance is itself accountable.

### *3.7. The Reversal Register*

The *Reversal Register* is a persistent log that records, for each decision or decision batch:

1. **State:** The authority state at the time of decision (HL, AL, CO, MO).
2. **Trigger:** The trigger class(es) that prompted entry into the current state.
3. **Thresholds:** The numeric thresholds in effect ($\theta\_D$, $\theta\_C$, $\theta\_R$, hysteresis width, timer duration).
4. **Actors:** Identifiers for the human operator(s) and AI system(s) involved.
5. **Explanations delivered:** The class and content of explanations presented.
6. **Actions taken:** Operator and system actions (approval, override, demand for elaboration, abort).
7. **Outcome (if known):** Decision result and any downstream consequences.

 doi:10.20944/preprints202603.0390.v1



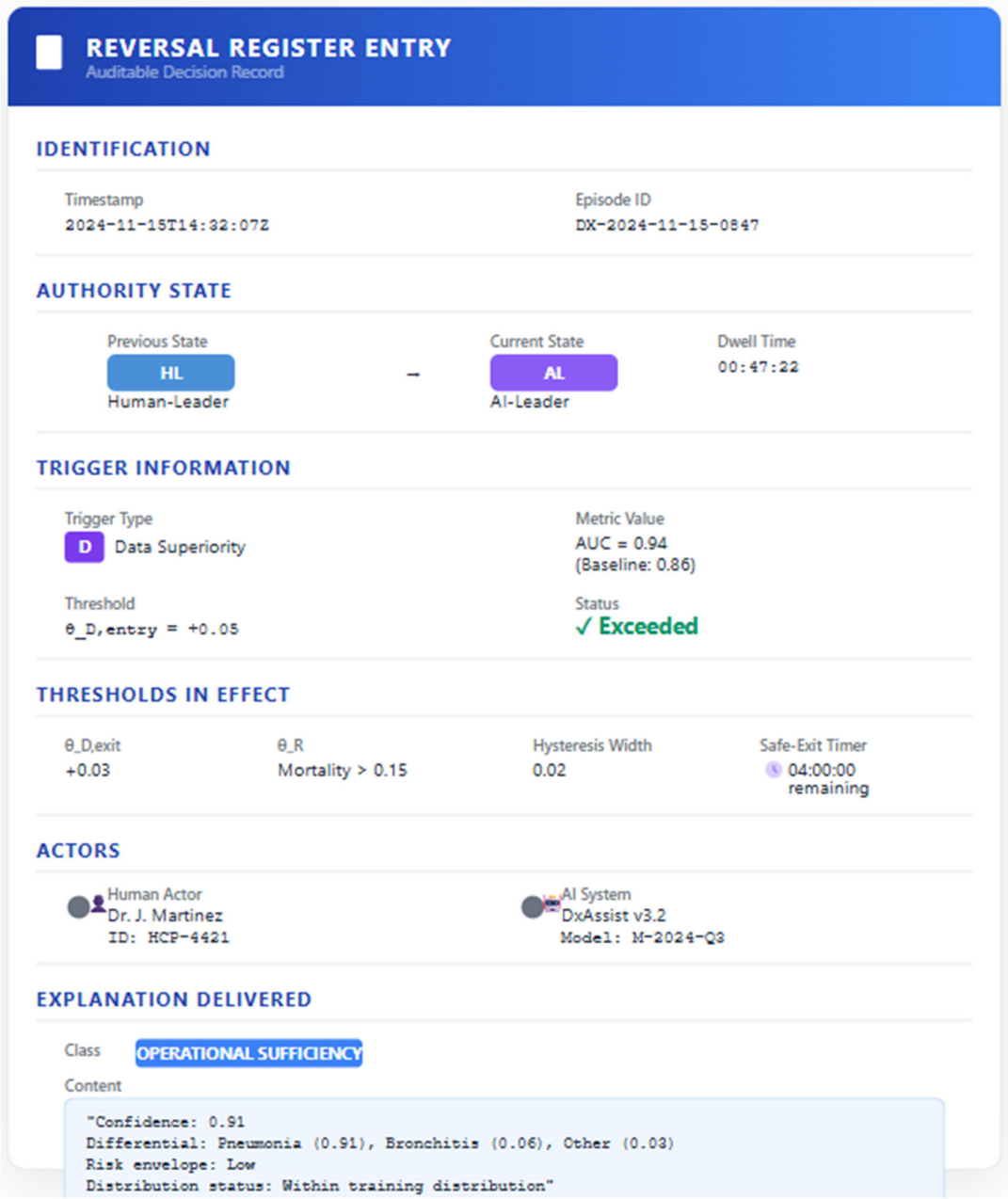


***Figure 3.*** *Reversal Register Entry Schema.*

The Reversal Register enables retrospective auditing, regulatory compliance, and organizational learning. By binding decisions to authority states, it operationalizes transitional accountability and provides the evidentiary basis for liability attribution.

### *3.8. Metrics and Telemetry*

To support empirical evaluation and continuous improvement, DAR specifies a telemetry stack comprising the following metrics:

| **Metric** | **Definition** | **Target Direction** |
|---|---|---|
| *Reversal latency* | Time (ms or decision cycles) from trigger detection to effective state change | Minimize |

 doi:10.20944/preprints202603.0390.v1



| Metric | Definition | Target Direction |
|---|---|---|
| *Hysteresis width* | Gap between entry and exit thresholds | Tune to balance stability and responsiveness |
| *Trust-whiplash index* | Incidence of user reliance shifts (over-trust ↔ under-trust) after reversals | Minimize |
| *Thrash rate* | Unwanted rapid state flips per 1,000 decisions | Near zero |
| *Phase-responsibility completion* | Percentage of decisions with complete Reversal Register entries | 100% |
| *Auditor-satisfaction score* | External-auditor rating of log completeness and traceability | Maximize |
| *Operator disorientation index* | Self-reported or behavioral measure of mode confusion following transitions | Minimize |

These metrics can be logged automatically (reversal latency, thrash rate, phase-responsibility completion) or collected through surveys and audits (trust-whiplash, auditor satisfaction, operator disorientation).

## 4. Propositions, Prioritization, and Methodological Pathways

DAR yields a set of falsifiable propositions linking authority configuration, transition mechanics, and system instrumentation to outcomes of interest. This section presents ten propositions, grouped thematically, with measurement constructs, prioritization rationale, and methodological considerations including threats to validity.

### *4.1. Proposition Prioritization*

The ten propositions differ in foundational importance and empirical tractability. **Foundational propositions** establish the core value of DAR; their falsification would undermine the framework. **Derivative propositions** refine the framework or explore boundary conditions; their falsification would prompt revision rather than rejection. **Tractability** refers to the feasibility of rigorous empirical testing given current methods and access.

| Proposition | Type | Tractability | Priority |
|---|---|---|---|
| P1 (Authority–Performance Fit) | Foundational | Moderate | High |
| P2 (Trigger Validity) | Foundational | Moderate | High |
| P3 (Hysteresis and Stability) | Derivative | High (simulation) | Medium |
| P4 (Safe-Exit Efficacy) | Derivative | High (experiment) | Medium |
| P5 (State-Contingent Explanations) | Foundational | High (experiment) | High |
| P6 (Cognitive-Load Moderation) | Derivative | High (experiment) | Low |
| P7 (Reversal Latency and Safety) | Derivative | Moderate | Medium |
| P8 (Override Friction and Error Detection) | Derivative | High (experiment) | Medium |
| P9 (Reversal Register and Legitimacy) | Derivative | Low (longitudinal) | Low |
| P10 (Transitional Accountability and Compliance) | Derivative | Low (archival) | Low |

 doi:10.20944/preprints202603.0390.v1



Researchers are encouraged to prioritize P1, P2, and P5 in initial empirical programs, as these establish the foundational claims of the framework.

### *4.2. Authority–Performance Fit*

**Proposition 1 (Authority–Performance Fit).** Decisions made under dynamically allocated authority outperform decisions made under static role allocations in terms of accuracy, efficiency, and risk-adjusted outcomes.

*Rationale.* If AI outperforms humans under certain conditions and humans outperform AI under others, a policy that allocates leadership to the superior party should dominate a policy that fixes leadership ex ante.

*Measurement constructs.* Decision accuracy (e.g., classification accuracy, prediction error), efficiency (time-to-decision, cost), risk-adjusted return (e.g., Sharpe ratio, harm-avoidance rate).

*Methodological pathway.* Randomized controlled trial or A/B test comparing DAR-governed workflows to static-HITL baselines, controlling for task complexity and operator experience.

*Threats to validity.* (a) The relevant counterfactual (what would have happened under static allocation) is not directly observable; randomization addresses this but may not capture real-world dynamics. (b) Ethical concerns may preclude randomization in high-stakes settings. Mitigations include simulation studies, stepped-wedge designs, or natural experiments exploiting exogenous variation in DAR adoption.

**Proposition 2 (Trigger Validity).** Transitions triggered by validated signals (data superiority, contextual-judgment proxies, risk metrics, ethics flags) yield greater performance gains than transitions triggered by arbitrary schedules or operator discretion alone.

*Rationale.* Validated triggers encode domain-relevant information; arbitrary schedules or unconstrained discretion may misalign authority with situational demands.

*Measurement constructs.* Trigger-valid positive rate (proportion of transitions followed by improved outcomes), false-trigger rate (proportion followed by degraded outcomes).

*Methodological pathway.* Archival analysis of Reversal Register logs linked to decision outcomes; comparison of trigger-based versus discretionary transitions.

*Threats to validity.* Selection effects—operators who invoke discretionary overrides may differ systematically from those who do not. Propensity-score matching or instrumental-variable designs can mitigate.

### *4.3. Stability and Responsiveness*

**Proposition 3 (Hysteresis and Stability).** Wider hysteresis bands reduce thrash rate but, beyond a threshold width, delay warranted reversals and degrade performance.

*Rationale.* Hysteresis trades off stability against responsiveness. Very narrow bands induce oscillation; very wide bands leave suboptimal authority states in place too long.

*Measurement constructs.* Thrash rate, reversal latency, decision accuracy.

*Methodological pathway.* Simulation studies varying hysteresis width; field pilots with randomized band settings.

*Threats to validity.* Simulation fidelity—results depend on accurate modeling of performance dynamics. Field pilots may face confounds from concurrent changes.

**Proposition 4 (Safe-Exit Efficacy).** Introducing safe-exit timers increases acceptance of AI-led states and reduces over-reliance without impairing operational efficiency.

*Rationale.* Operators accept bounded autonomy when they know human review is guaranteed within a predictable interval.

*Measurement constructs.* Operator acceptance (survey scale), over-reliance rate (proportion of cases where AI errors went undetected), efficiency (throughput per unit time).

*Methodological pathway.* Between-subjects experiment comparing conditions with and without safe-exit timers.

 doi:10.20944/preprints202603.0390.v1



*Threats to validity.* Demand characteristics—operators aware of the timer condition may behave differently. Behavioral measures (error detection) can supplement self-report.

### *4.4. Explainability and Trust Calibration*

**Proposition 5 (State-Contingent Explanations).** Explanations tailored to authority state improve trust calibration and auditor satisfaction relative to generic explanations.

*Rationale.* Explanation needs differ by state: summative in HL, operational-sufficiency in AL, diagnostic in CO, justificatory in MO.

*Measurement constructs.* Trust-calibration index (alignment between operator reliance and AI reliability), auditor-satisfaction score, explanation-completeness rating.

*Methodological pathway.* Within-subjects experiment manipulating explanation type across simulated states; field study measuring audit outcomes.

*Threats to validity.* Carryover effects in within-subjects designs. Between-subjects replication can assess robustness.

**Proposition 6 (Cognitive-Load Moderation).** The positive effect of state-contingent explanations on trust calibration is stronger when cognitive load is moderate than when load is very low (floor effect) or very high (overload).

*Rationale.* Under minimal load, operators can process any explanation format; under extreme load, even tailored explanations may be ignored.

*Measurement constructs.* Cognitive load (secondary-task performance, self-report), trust-calibration index.

*Methodological pathway.* Factorial experiment crossing explanation type with induced cognitive-load conditions.

*Threats to validity.* Induced load may not replicate real-world conditions. Field observation of naturally varying load can supplement lab findings.

### *4.5. Safety and Latency*

**Proposition 7 (Reversal Latency and Safety).** Shorter reversal latency—time from valid trigger to effective state change—reduces near-miss incidents and tail-risk events in high-stakes workflows.

*Rationale.* Delayed handovers prolong exposure to suboptimal authority configurations.

*Measurement constructs.* Reversal latency, near-miss rate, tail-risk incidence (e.g., losses beyond a threshold).

*Methodological pathway.* Event-study analysis of historical workflows with varying latencies; simulation with injected delays.

*Threats to validity.* Endogeneity—workflows with shorter latency may differ in other ways. Instrumental-variable or regression-discontinuity designs can address.

**Proposition 8 (Override Friction and Error Detection).** Reducing procedural friction for MO invocation increases the rate at which operators detect and override erroneous AI actions.

*Rationale.* High friction—e.g., multi-step confirmations, managerial sign-off—discourages override even when warranted.

*Measurement constructs.* Friction score (steps and time required for override), error-detection rate, false-override rate.

*Methodological pathway.* Field experiment randomizing friction levels across matched units.

*Threats to validity.* Reducing friction may increase false overrides (Type I errors). Analysis should examine both detection and false-override rates.

### *4.6. Accountability and Legitimacy*

**Proposition 9 (Reversal Register and Legitimacy).** Organizations that maintain a complete Reversal Register achieve higher stakeholder-trust ratings, better audit outcomes, and stronger legal defensibility than organizations that rely on informal documentation.

 doi:10.20944/preprints202603.0390.v1



*Rationale.* The Reversal Register creates an evidentiary trail that satisfies regulatory expectations and supports liability attribution.

*Measurement constructs.* Stakeholder-trust survey, audit rating, legal-case outcomes (where available).

*Methodological pathway.* Comparative case study of organizations with and without register practices; archival analysis of regulatory filings.

*Threats to validity.* Legal defensibility is difficult to isolate—litigation outcomes depend on many factors. Audit ratings and stakeholder trust are more tractable proxies.

**Proposition 10 (Transitional Accountability and Compliance).** Clear assignment of accountability to the prevailing authority state reduces regulatory sanctions and improves compliance-audit scores.

*Rationale.* Regulators penalize ambiguity; clear accountability reduces interpretive disputes.

*Measurement constructs.* Sanction frequency, compliance-audit score.

*Methodological pathway.* Panel analysis of regulated entities adopting DAR versus non-adopters.

*Threats to validity.* Selection effects—entities adopting DAR may be more compliance-oriented ex ante. Difference-in-differences or synthetic-control methods can mitigate.

### *4.7. Research-Ethics Considerations*

Several propositions involve interventions in high-stakes settings (healthcare, public administration). Empirical validation of DAR will require careful ethical review, including:

- Informed consent for operators participating in experiments.
- Equipoise considerations—randomization is appropriate only when there is genuine uncertainty about which condition is superior.
- Monitoring and stopping rules—experiments should include interim analyses to halt if one condition proves clearly harmful.
- Privacy protections for data in Reversal Registers.

Researchers should consult institutional review boards and domain-specific ethics guidelines before conducting field research.

### *4.8. Minimal Telemetry Specification*

To test the foregoing propositions, systems should emit a minimal telemetry set comprising:

- Current state (HL, AL, CO, MO)
- Trigger type(s) active
- Entry/exit thresholds in effect
- Dwell time in current state
- Safe-exit timer status (time remaining)
- Explanation class delivered
- Actor action sequence (approve, override, request elaboration, abort)

From these primitives, derived metrics—reversal latency, hysteresis width, trust-whiplash index, thrash rate, phase-responsibility completion—can be computed.

## 5. Sector-Specific Implementation Guidance

This section provides in-depth implementation guidance for two sectors: healthcare and public administration. By focusing on two sectors rather than four, the paper offers greater depth, engaging with sector-specific literatures, governance structures, and regulatory frameworks. The goal is illustrative rather than exhaustive; readers are directed to sector-specific literatures for further elaboration.

### *5.1. Healthcare*

#### 5.1.1. Context and Existing Governance Structures

 doi:10.20944/preprints202603.0390.v1



Clinical decision-support systems (CDSS) promise improved diagnostic accuracy and treatment selection, yet clinician uptake remains uneven, and liability questions persist (Topol, 2019). The Therac-25 catastrophe remains a cautionary reminder that absent explicit handover protocols, software-driven medical systems can fail catastrophically (Leveson & Turner, 1993).

Healthcare organizations operate within layered governance structures: professional licensure (physicians, nurses), institutional review (hospital ethics committees, quality-improvement committees), regulatory oversight (FDA, CMS), and legal liability (malpractice doctrine). Implementation of DAR must align with these structures.

Research on CDSS implementation barriers identifies several recurrent challenges: alert fatigue, workflow disruption, lack of trust, and unclear liability (Sittig & Singh, 2010). DAR addresses several of these by (a) limiting alerts to state-relevant triggers rather than generic warnings, (b) aligning transitions with workflow phases, (c) providing state-contingent explanations that support calibrated trust, and (d) documenting authority states to clarify liability.

*5.1.2. DAR Implementation*

**States:**

- *HL* for complex, multi-morbid, or preference-sensitive cases, where clinician judgment is essential.
- *AL* for well-characterized screening tasks (e.g., diabetic-retinopathy grading) where AI meets regulatory-performance thresholds and case characteristics fall within the AI's validated operating envelope.
- *CO* for multidisciplinary tumor boards or shared decision-making encounters, where a documented merge protocol combines clinician and AI inputs.
- *MO* for patient-safety alerts (e.g., contraindication flags, drug–drug interaction warnings) or when cases fall outside the AI's validated envelope.

**Triggers:**

- *Data superiority:* AI accuracy exceeding physician baseline on a validated, representative test set by a pre-specified margin (e.g., +5% sensitivity). Validation should use local data reflecting the institution's case mix.
- *Contextual judgment:* Patient-preference signals (e.g., documented values, advance directives), novel symptom constellations, social determinants flagged by intake screening, or model uncertainty exceeding threshold.
- *Risk threshold:* Mortality-risk score or complication-probability exceeding threshold; severity of potential adverse outcome.
- *Ethics override:* Informed-consent requirements, off-label prescribing, resource-allocation dilemmas (e.g., ICU triage), protected-class considerations.

**Guardrails:**

- *Safe-exit timers* keyed to clinical-workflow stages (e.g., 24-hour review for admitted patients, end-of-shift review for outpatient clinics). Timers ensure that even if the AI leads, a clinician reviews the case within a defined interval.
- *Hysteresis* on accuracy thresholds to accommodate inter-shift variability and prevent oscillation due to random fluctuations in case mix.

**Trigger Governance:**

- A multidisciplinary committee (physicians, nurses, informaticists, ethicists, patient representatives) calibrates thresholds, adjudicates disputes, and reviews performance quarterly.
- Committee decisions are documented and accessible to the hospital's quality-improvement and compliance functions.

**Reversal Register:**

- Linked to electronic health record (EHR) audit logs, ensuring that each diagnostic or treatment decision is associated with the prevailing authority state.

 doi:10.20944/preprints202603.0390.v1



- Entries accessible for malpractice defense, quality-improvement review, and regulatory inspection.
- Privacy protections ensure that patient-identifiable information is handled in accordance with HIPAA and institutional policies.

**Integration with Existing Governance:**

- *Morbidity and Mortality (M&M) Conferences:* Reversal Register data can inform M&M reviews by identifying cases where authority transitions may have contributed to adverse outcomes.
- *Quality-Improvement Committees:* Aggregate telemetry (thrash rate, reversal latency, trigger distributions) supports continuous improvement.
- *Liability Allocation:* Malpractice doctrine generally holds the treating clinician responsible. DAR clarifies that in HL, the clinician is unambiguously accountable; in AL, the clinician as designated monitor is accountable for failures to override detectable errors; in CO, accountability follows the documented merge protocol; in MO, the invoking party is accountable for the halt. These clarifications align with, rather than displace, existing doctrine.

#### 5.1.3. Expected Benefits

- Reduced diagnostic delay where AI leads appropriately within its validated envelope.
- Preserved clinician authority for complex or values-laden decisions.
- Reduced alert fatigue by limiting alerts to state-relevant triggers.
- Clear documentation supporting liability attribution and quality improvement.
- Enhanced patient trust through transparent authority configurations (see Section 6.4 on affected publics).

#### 5.1.4. Challenges and Mitigations

- *Clinician resistance:* Some clinicians may resist ceding authority to AI. Change management should emphasize that DAR preserves clinician authority for complex cases and that AI leadership applies only within validated envelopes with safe-exit guarantees.
- *Technical integration:* Integrating DAR with existing EHR systems may require significant development effort. Phased rollout—starting with a single workflow (e.g., retinal screening) before expanding—can manage complexity.
- *Regulatory uncertainty:* Regulatory guidance on CDSS liability is evolving. Organizations should engage with FDA and CMS to ensure that DAR implementations align with emerging requirements.

### *5.2. Public Administration*

#### 5.2.1. Context and Existing Governance Structures

Government agencies deploy predictive analytics for child-welfare screening, recidivism assessment, benefit eligibility, and tax-fraud detection. Critics have documented due-process concerns, opacity, and discriminatory impacts (Eubanks, 2018). Regulatory guidance increasingly mandates human review but lacks operational specificity.

Public-administration contexts are governed by administrative-law doctrines, including procedural due process, reasoned decision-making requirements, and rights to notice and appeal. Fountain (2001) documents extensive evidence of resistance to technology adoption in government, driven by institutional inertia, risk aversion, and power dynamics. DAR implementation must navigate these constraints.

#### 5.2.2. DAR Implementation

**States:**

 doi:10.20944/preprints202603.0390.v1



- *HL* for final determinations affecting fundamental rights (e.g., benefit denial, child removal, parole decisions), where due-process requirements mandate substantive human judgment.
- *AL* for low-stakes triage or scheduling (e.g., prioritizing case review order, scheduling appointments), where AI leadership improves efficiency without implicating fundamental rights.
- *CO* for intermediate-risk cases (e.g., moderate risk scores in child-welfare screening), where caseworker and model inputs are combined via a structured deliberation protocol.
- *MO* for civil-liberties alerts (e.g., protected-class flags, anomalous-score patterns suggesting data error), procedural-fairness requirements, or when the affected party contests the determination.

**Triggers:**

- *Data superiority:* Model accuracy validated on representative local data, with attention to subgroup performance (accuracy disaggregated by race, gender, geography).
- *Contextual judgment:* Case complexity flags (e.g., multiple interacting risk factors), out-of-distribution indicators, caseworker annotation of situational complexity.
- *Risk threshold:* Severity scores exceeding threshold (e.g., child safety risk exceeding 0.8), potential for irreversible harm (e.g., child removal).
- *Ethics override:* Protected-class sensitivity (disparate-impact flags), procedural-fairness requirements (e.g., right to hearing), whistleblower or fraud-allegation context.

**Guardrails:**

- *Safe-exit timers* aligned with statutory deadlines (e.g., 30-day benefit determination, 48-hour child-safety response). Timers ensure that AI-led triage does not circumvent statutory requirements.
- *Hysteresis* to prevent rapid oscillation in risk classifications, which could confuse caseworkers and undermine trust.

**Trigger Governance:**

- A committee including agency leadership, frontline workers, legal counsel, civil-liberties advocates, and community representatives calibrates thresholds and adjudicates disputes.
- Committee composition should reflect the diversity of affected publics to ensure that threshold-setting is not captured by any single interest.
- Decisions are documented and subject to Freedom of Information Act (FOIA) requests.

**Reversal Register:**

- Maintained as administrative record, subject to FOIA and judicial review.
- Entries retained for the duration required by records-retention schedules.
- Register serves as evidentiary basis for administrative appeals: affected parties can request documentation of the authority state, triggers, and explanations that governed their case.

**Integration with Existing Governance:**

- *Administrative-Law Doctrine:* Reasoned decision-making requirements (e.g., under the Administrative Procedure Act) demand that agencies explain the basis for determinations. The Reversal Register provides structured documentation that supports compliance.
- *Appellate Review:* When determinations are challenged, courts can examine the Reversal Register to assess whether the agency appropriately allocated authority and responded to triggers.
- *Legislative Oversight:* Aggregate telemetry (state distributions, trigger frequencies, outcome disparities) can inform legislative hearings on algorithmic governance.

5.2.3. Expected Benefits

- Efficiency gains in routine processing (triage, scheduling) without sacrificing due process for consequential decisions.
- Enhanced due-process protections through documented deliberation and clear authority assignment.
- Defensible administrative record supporting judicial review and legislative oversight.
- Reduced disparate impact through disaggregated monitoring and protected-class triggers.

5.2.4. Challenges and Mitigations

 doi:10.20944/preprints202603.0390.v1



- *Political resistance:* Threshold-setting allocates power; managers, caseworkers, and external stakeholders may resist changes that affect their authority. Inclusive governance (diverse committee composition) and transparent documentation can build buy-in.
- *Capacity constraints:* Frontline workers in under-resourced agencies may lack time for additional deliberation. DAR should be designed to reduce, not increase, overall workload by concentrating human effort on high-stakes cases.
- *Transparency demands:* Public-administration contexts face heightened transparency expectations. Organizations should consider proactive disclosure of aggregate telemetry (e.g., annual reports on state distributions and outcomes) to build public trust.

### *5.3. Additional Sectors: Summary Guidance*

While the paper focuses in depth on healthcare and public administration, DAR is applicable to other sectors. Brief summary guidance is offered below; readers are directed to sector-specific literatures for further elaboration.

**Financial Services:**

- *States:* HL for strategy and limit-setting; AL for execution within limits; CO for novel instruments; MO for circuit-breaker events.
- *Key considerations:* Integration with existing risk-management frameworks; alignment with prudential-regulation requirements; real-time telemetry for high-frequency environments.

**Human-Resource Management:**

- *States:* HL for final hiring/promotion decisions; AL for initial résumé screening; CO for calibration sessions; MO for adverse-impact alerts.
- *Key considerations:* Compliance with employment-discrimination law (e.g., Title VII, EEOC guidance); attention to candidate experience and dignity; integration with applicant-tracking systems.

 doi:10.20944/preprints202603.0390.v1



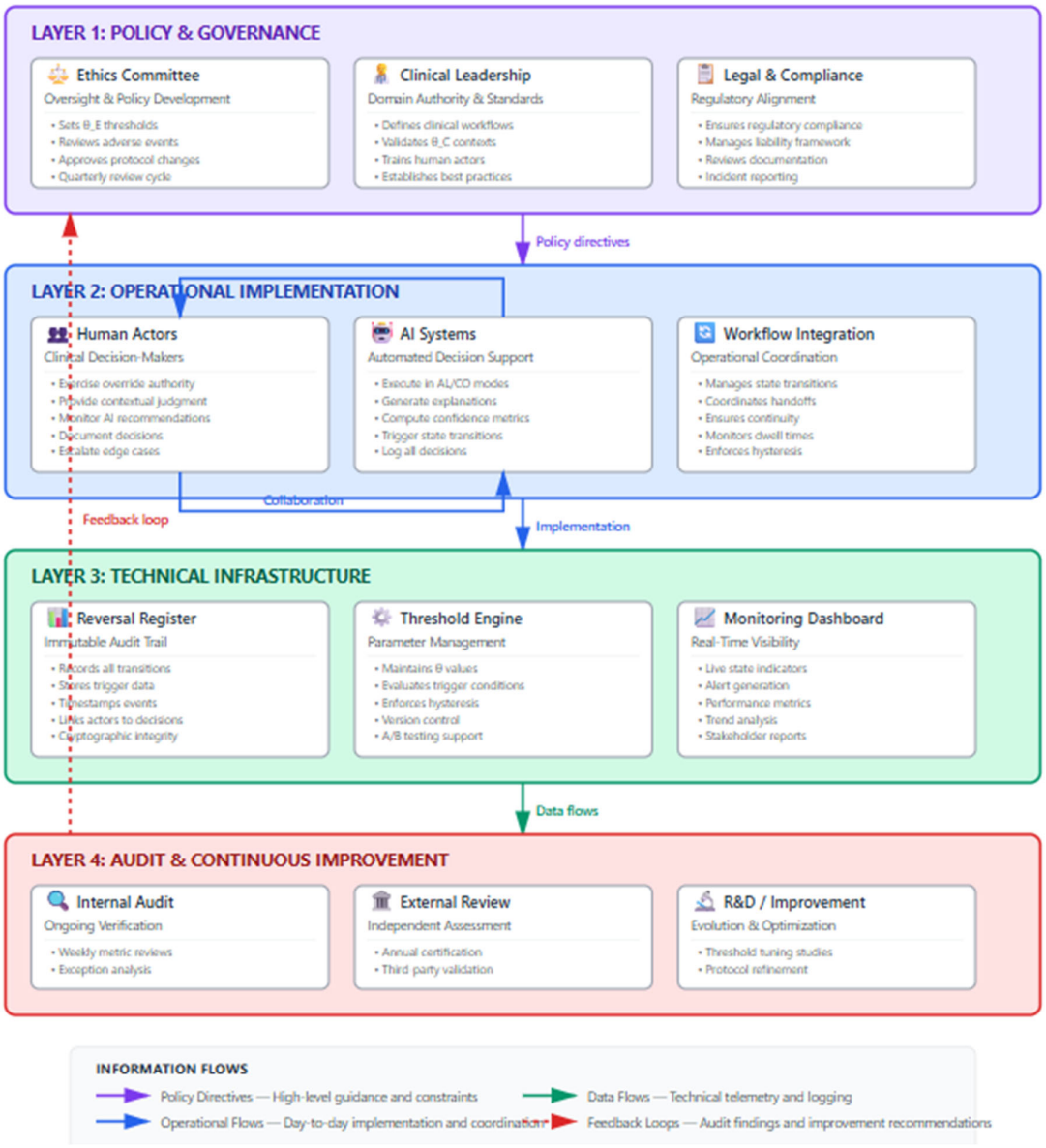


**Figure 4.** Governance Structure Diagram.

## 6. Discussion

### *6.1. Theoretical Contributions*

DAR advances human–AI collaboration theory along several dimensions.

**Transition logic.** DAR introduces transition logic as a first-class construct, specifying when and why authority should shift. By deriving the state taxonomy from the intersection of initiative and enactment authority, DAR provides a principled foundation rather than an ad hoc classification. The framework integrates insights from the adaptive-automation tradition (Kaber & Endsley, 2004; Sheridan & Verplank, 1978) while extending that tradition to address organizational accountability contexts.

**Multidimensional accountability.** By engaging Bovens◉(2007) accountability framework and Wieringa◉ (2020) extension to algorithmic contexts, DAR theorizes transitional accountability along dimensions of answerability, responsibility, and liability. The Reversal Register operationalizes each

 doi:10.20944/preprints202603.0390.v1



dimension, providing structured documentation for retrospective justification, duty assignment, and liability attribution.

**Integration across literatures.** DAR bridges disparate literatures—adaptive automation, shared leadership, trust calibration, explainable AI, algorithmic governance—into a unified framework. This integration enables cumulative research by providing common constructs, metrics, and propositions.

### *6.2. Practical Implications*

For practitioners, DAR offers a governance-ready framework. The Authority-State Playbook predefines roles, triggers, thresholds, and explanation strategies, reducing ambiguity and enabling consistent implementation across units. Safe-exit timers operationalize meaningful human oversight, satisfying regulatory expectations without requiring continuous vigilance. The Reversal Register provides an auditable trail that supports internal review, external audit, and legal defense.

Implementation, however, is not trivial. Organizations must invest in threshold calibration, operator training, and technical infrastructure for logging and telemetry. Change-management challenges may arise as roles become more fluid and as operators accustom themselves to conditional authority. Pilot programs, phased rollouts, and continuous-improvement cycles are recommended to surface and address these challenges incrementally.

### *6.3. The Human Experience of Handovers*

A recurrent finding in the human-factors literature is that transitions between automation modes can produce disorientation, confusion, and error (Sarter & Woods, 1995). DAR must attend not only to the formal structure of handovers but also to the *phenomenology* of handovers—how operators experience, interpret, and respond to authority transitions.

Several design considerations can support positive operator experience:

- *Clear signaling:* Interface elements should unambiguously indicate the current authority state (e.g., color coding, status banners, auditory cues).
- *Transition previews:* When a trigger is approaching but has not yet fired, operators can be alerted to the impending transition, reducing surprise.
- *Contestability:* Operators should have a low-friction mechanism to contest a transition they believe is unwarranted. The Trigger Governance Committee can review contested transitions retrospectively.
- *Training:* Operators should receive training on the logic of the state taxonomy, the meaning of triggers, and the expectations for their role in each state.
- *Workload management:* Frequent transitions can impose cognitive load. Hysteresis and safe-exit timers should be calibrated to balance responsiveness against operator burden.

Empirical research (Proposition 6, Proposition 8) can inform the design of these features by identifying conditions under which transitions support versus undermine operator performance.

### *6.4. Political Dimensions of DAR Implementation*

DAR is presented as a rational-technical framework, but its implementation inevitably encounters organizational politics. Threshold-setting allocates power: lower data-superiority thresholds empower AI and those who champion it; higher thresholds empower human operators and their advocates. Managers may resist ceding authority to AI; operators may resist ceding authority to automated triggers; AI developers may resist constraints on system autonomy.

Acknowledging these dynamics, DAR incorporates governance mechanisms—the Trigger Governance Committee, documented threshold decisions, periodic review—that render political contestation visible and manageable. By including diverse stakeholders in governance, organizations can negotiate thresholds in a transparent forum rather than allowing power to be exercised covertly through technical design choices.

 doi:10.20944/preprints202603.0390.v1



This approach resonates with Fountain's (2001) emphasis on institutional factors in technology adoption. DAR does not assume that rational-technical design will overcome political resistance; rather, it provides structures that channel contestation into accountable deliberation.

### *6.5. Affected Publics*

DAR is framed primarily from the perspective of organizations and regulators, but consequential AI-assisted decisions affect members of the public—patients, job applicants, welfare recipients—who have legitimate interests in how authority is allocated.

Several considerations arise:

- *Transparency to affected parties:* Should individuals be informed when their case was governed by AI leadership versus human leadership? In principle, the Reversal Register provides this information; in practice, disclosing detailed authority configurations may overwhelm or confuse non-expert audiences. Organizations should consider summary disclosures (e.g., "This decision was reviewed by a physician with assistance from an AI diagnostic tool") that convey the essential authority structure without technical detail.
- *Contestability by affected parties:* Affected individuals should have mechanisms to contest decisions, including access to information about the authority configuration that governed their case. In public-administration contexts, this aligns with due-process requirements; in other contexts, organizations may adopt analogous practices voluntarily.
- *Participation in governance:* Where feasible, affected publics—or their representatives—should participate in Trigger Governance Committees. Community representatives can ensure that threshold-setting reflects public values, not only organizational interests.

These considerations extend DAR's accountability logic from intra-organizational dynamics to the broader ecosystem of affected stakeholders.

### *6.6. Tensions and Trade-offs*

DAR does not resolve all tensions inherent in human–AI collaboration; rather, it surfaces and structures them. Three tensions warrant explicit acknowledgment:

**Efficiency vs. legitimacy.** Frequent human checkpoints (via safe-exit timers) satisfy accountability requirements but may degrade throughput and operator satisfaction. Organizations must calibrate timer durations to balance these demands, recognizing that the optimal balance may differ across workflows and risk levels.

**Stability vs. responsiveness.** Wider hysteresis bands reduce thrash but may delay warranted reversals. There is no universally optimal width; organizations should use simulation and empirical feedback to identify appropriate settings for their context.

**Operator autonomy vs. system control.** If the system determines when authority shifts (based on trigger thresholds), operators may experience reduced autonomy. Interface design (contestability mechanisms, transition previews) can mitigate this concern, but some tension is inherent in any structured governance framework.

Acknowledging these trade-offs forthrightly enhances the paper's credibility and practical value. DAR provides a framework within which trade-offs can be negotiated, not a formula that eliminates them.

### *6.7. Regulatory Alignment*

DAR aligns with emerging regulatory frameworks, notably the EU AI Act, which mandates human oversight, transparency, and documentation for high-risk AI systems (Regulation (EU) 2024/1689). The Reversal Register satisfies documentation requirements; state-contingent explanations address transparency expectations; safe-exit timers and MO provisions operationalize human-oversight mandates.

 doi:10.20944/preprints202603.0390.v1



Beyond Europe, regulatory developments in the United States (e.g., proposed algorithmic-accountability legislation, sector-specific guidance from financial and health regulators) and in other jurisdictions reflect a global trajectory toward mandated human oversight. DAR's modular design—states, triggers, guardrails, logs—allows organizations to adapt to varying regulatory requirements by adjusting thresholds and timer durations without redesigning the underlying architecture.

### *6.8. Limitations and Boundary Conditions*

DAR is subject to several limitations:

**Domain-specific calibration.** Trigger operationalization requires domain-specific calibration; generic thresholds may misalign with local performance distributions. Organizations must invest in validation studies before deployment.

**Parameter tuning.** Hysteresis and timer parameters demand empirical tuning; poorly chosen values can induce instability or rigidity. Simulation and pilot studies are essential.

**Real-time observability.** The framework assumes that transitions can be detected and executed in near-real-time; workflows with high latency or low observability may require adaptation.

**Designated human owner.** DAR presupposes that a designated human owner can be identified for each workflow; highly distributed or anonymous decision contexts may complicate accountability attribution.

**Generalizability to low-stakes and creative tasks.** DAR is designed for high-stakes workflows where authority and accountability are salient. In low-stakes or exploratory/creative tasks, authority may be more fluid and less amenable to discrete-state modeling. Future work might develop "soft DAR" variants with more negotiated, less formalized transitions for such contexts.

**Cultural and jurisdictional variation.** Trust calibration, accountability expectations, and regulatory frameworks vary across cultures and legal systems. Comparative research is needed to identify moderators and inform localized implementations.

### *6.9. Future Research Directions*

Several avenues merit further investigation:

1. **Empirical validation.** Field experiments and longitudinal studies are needed to test the propositions derived in Section 4 across diverse domains. Priority should be given to foundational propositions (P1, P2, P5).
2. **Parameter optimization.** Machine-learning techniques (e.g., reinforcement learning, Bayesian optimization) could automate hysteresis and timer tuning based on historical performance data, subject to safety constraints.
3. **Multi-agent extensions.** DAR currently models a single AI system and a single human party; extension to multi-AI and multi-human configurations is warranted, particularly for team-based workflows.
4. **Explanation-generation pipelines.** Research should develop methods for generating state-contingent explanations automatically, drawing on large language models and domain ontologies.
5. **Cross-cultural and cross-jurisdictional variation.** Comparative research can identify moderators and inform localized implementations.
6. **Affected-public engagement.** Empirical research on how affected publics perceive and respond to different authority configurations can inform transparency and contestability design.
7. **Soft-DAR variants.** For creative, exploratory, or low-stakes tasks, variants with softer, more negotiated transitions may be appropriate; conceptual and empirical work is needed to develop such variants.

## 7. Conclusions

Human–AI collaboration is not a static allocation problem but a dynamic coordination challenge. The Dynamic Authority Reversal (DAR) framework addresses this challenge by modeling intra-

 doi:10.20944/preprints202603.0390.v1



episode authority transitions across four states—Human-Leader/AI-Follower, AI-Leader/Human-Follower, Co-Leadership, and Mutual Override—governed by principled triggers, stabilized by hysteresis and safe-exit timers, and documented through the Reversal Register. DAR operationalizes the abstract expectation of "human oversight" into measurable, auditable practice, bridging micro-level trust calibration with macro-level legitimacy and compliance.

For researchers, DAR offers ten falsifiable propositions—prioritized by foundational importance and tractability—a telemetry specification, and a minimal-instrumentation stack that support cumulative empirical inquiry. For practitioners, it furnishes governance-ready artifacts—Authority-State Playbooks, state-contingent explanation templates, Reversal Registers—that translate theory into deployable systems. For regulators, it provides an auditable architecture that satisfies emerging legal requirements while preserving flexibility for domain-specific adaptation. For affected publics, it offers structured transparency and contestability mechanisms that support trust and accountability.

The Therac-25 catastrophe of 1985–1987 revealed what happens when handovers fail—when authority is ceded to machines without meaningful reversion mechanisms, and when operators lack the signals and protocols needed to intervene. DAR ensures that handovers—between human and AI, across states, over time—are not afterthoughts but first-order design considerations. By making authority transitions visible, principled, and accountable, DAR advances both the effectiveness and the legitimacy of human–AI collaboration in high-stakes workflows.

## References


Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, *13*(4), 447–468.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21.

D'Innocenzo, L., Mathieu, J. E., & Kukenberger, M. R. (2016). A meta-analysis of different forms of shared leadership–team performance relations. *Journal of Management*, *42*(7), 1964–1991.

Davenport, T. H., & Kirby, J. (2016). *Only humans need apply: Winners and losers in the age of smart machines*. Harper Business.

DeRue, D. S., & Ashford, S. J. (2010). Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of Management Review*, *35*(4), 627–647.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*.

Fountain, J. E. (2001). *Building the virtual state: Information technology and institutional change*. Brookings Institution Press.

Goodrich, M. A., & Schultz, A. C. (2007). Human–robot interaction: A survey. *Foundations and Trends in Human–Computer Interaction*, *1*(3), 203–275.

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99.

Gronn, P. (2002). Distributed leadership as a unit of analysis. *The Leadership Quarterly*, *13*(4), 423–451.

 doi:10.20944/preprints202603.0390.v1



Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Leveson, N. G., & Turner, C. S. (1993). An investigation of the Therac-25 accidents. *Computer*, *26*(7), 18–41.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507.

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *30*(3), 286–297.

Pearce, C. L., & Conger, J. A. (2003). *Shared leadership: Reframing the hows and whys of leadership*. SAGE Publications.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), 192–210.

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 the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5–19.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical Report). MIT Man-Machine Systems Laboratory.

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, *19*(Suppl 3), i68–i74.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

 doi:10.20944/preprints202603.0390.v1



**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.