IX. International Applied Social Sciences Congress - C-iasoS 2025
Sapienza University of Rome, Italy, 13-15 October 2025

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

Victor Frimpong[1*], Charles Tawk[2], Agim Mamuti[3]

[1]Management Department, SBS Swiss Business School, Zurich, Switzerland
[2]Meirc Training and Consulting: Dubai, UAE
[3]Faculty of Economics, Technology and Innovation, Western Balkans University, Tirana, Albania
[*]*Corresponding author:* v.frimpong@research.sbs.edu

Abstract

Introduction: Human–AI collaboration is reconceptualized as a temporal authority problem: leadership shifts within a single decision episode across four states—Human Leader→AI Follower, AI Leader→Human Follower, Co-Leadership, and Mutual Override—triggered by data superiority, contextual judgment needs, risk thresholds, and ethics overrides.
Aim: To formalize the Dynamic Authority Reversal (DAR) framework, calibrating trust and aligning accountability with the active leadership state, and to translate it into testable propositions and an operational playbook.
Method: Conceptual synthesis of leader–follower, distributed leadership, and HITL literatures; state-and-transition architecture; derivation of ten propositions with measurement constructs (reversal latency, hysteresis, "trust whiplash") and a cross-sector implementation roadmap.
Findings: DAR surfaces and addresses five debates—decision ownership during AI-led phases, autonomy–acceptance trade-offs (mitigated by safe-exit guarantees), optimal reversal tempo, state-contingent explanations, and transitional accountability via role-state logging. It specifies triggers, guardrails, telemetry, and KPIs to make handovers auditable and performant.
Conclusion: By keeping humans ultimately responsible while enabling reversible AI leadership with measurable handovers, DAR improves effectiveness, legitimacy, and compliance in high-stakes workflows across finance, healthcare, public administration, and HR
Originality and value: DAR moves beyond static augmentation/autonomy/HITL categories by modeling intra-episode handovers and coupling theory to buildable instruments—Authority-State Playbooks, safe-exit timers, state-contingent XAI, and a Reversal Register—plus falsifiable propositions for cumulative research.

Keywords: human–AI collaboration; authority reversal, transitional accountability
Jel Codes: M54, O33, D83

1. Introduction

On May 6, 2010, U.S. financial markets experienced a significant event known as the "Flash Crash," during which a trillion-dollar decline occurred due to algorithmic trading until human regulators intervened to restore order. This incident underscores the potential for temporary shifts in authority to artificial intelligence until associated risks necessitate a return to human oversight.

Human–AI collaboration is typically viewed as a division of tasks, with machines performing analyses and humans making decisions. However, this viewpoint overlooks the dynamic nature of authority in real-world workflows, where leadership may change several times before a final decision is reached. In unstable settings characterized by rapidly evolving signals and changing risk parameters, effective leadership must be flexible and responsive.

This paper contends that traditional categories—augmentation, autonomy, and human-in-the-loop (HITL)—are no longer relevant. These classifications enforce inflexible roles on both humans and AI, primarily assessing success in terms of accuracy or utility. An essential element that is overlooked is a theory of transition logic, which involves understanding who takes the lead at various times, the reasons for leadership changes, and how accountability shifts during these transitions. Without this framework, organizations risk either overemphasizing human control, which can lead to suboptimal performance, or placing too much trust in AI, potentially causing issues connected to legitimacy,

safety, and compliance.

We introduce the Dynamic Authority Reversal (DAR) framework. DAR views leadership as a state machine comprising four states: HL→AF (humans in charge while AI follows), AL→HF (AI in charge while humans follow), CO (shared leadership), and MO (mutual override, allowing either party to stop or reverse the process). The transitions between these states are shaped by four main elements: data dominance (when AI surpasses a specific predictive benchmark), contextual decision requirements (human factors such as implicit knowledge), risk boundaries (when risks exceed established limits), and ethical constraints (normative constraints based on statutes or regulations). Safe-exit guarantees and hysteresis ensure a smooth transition, avoiding sudden changes. DAR presents three main contributions. First, it posits that altering leadership based on real-time information, advantages, or ethical considerations yields more effective outcomes than adhering to fixed roles. Second, it emphasizes the importance of legitimacy through transitional accountability, which connects decisions to the prevailing leadership at the time. Third, it underscores the need to implement accessible, practical tools (such as timers, registers, thresholds, and state-aware XAI) to operationalize transitions, rather than relying solely on "human oversight."

2. Literature Review

2.1 Static roles and handover logic

Research on augmentation and human-in-the-loop (HITL) typically assigns fixed roles to both humans and AI, offering design principles but lacking a definitive process for adjusting control during interactions (Amershi et al., 2019). Although research on autonomy emphasizes reliability and performance, it typically does not specify when humans should reassert control as situations change (Rahwan et al., 2019). Leadership theories, including distributed and shared leadership, acknowledge the participation of various agents but lack clear definitions of the signals for transitions or the associated guidelines (Wu, Cormican, & Chen, 2020; Chen & Zhang, 2022; Maritsa et al., 2022).

DAR fills this void by outlining clear states and transitions that facilitate consistent and verifiable shifts in authority. It establishes a transition logic that connects changes in authority to measurable triggers (like data superiority, contextual judgment, and ethical considerations) and guidelines (including hysteresis and safe exits). This approach aligns with calls to examine machine behavior in context, rather than in isolation (Rahwan et al., 2019).

2.2 Performance and legibility

Explainable AI (XAI) has advanced, yet it often assumes that decision boundaries are fixed. Studies indicate that heightened transparency does not automatically improve the quality of human-AI cooperation or help users recognize mistakes; instead, it may overwhelm users (Miller, 2019; Poursabzi-Sangdeh et al., 2021). The effectiveness of explanations in teamwork can either support or obstruct collaboration, depending on the nature of the task and the user's cognitive state (Bansal et al., 2020; Buçinca, Malaya, & Gajos, 2021). A growing focus in HCI is on creating user-centered, task-specific explanations that directly address users' questions (Liao et al., 2020a, 2020b) and on evaluating trust calibration, rather than just “trust” (Naiseh et al., 2023; Bollaert et al., 2024).

For DAR, explanations vary depending on the state of leadership. DAR supports state-contingent XAI: operational sufficiency in AL→HF (e.g., confidence, alternatives, risk envelope); diagnostic support in HL→AF; negotiation-grade rationale in CO; and override justification in MO. This approach treats explanations as tools for specific phases rather than generic documentation.

2.3 Micro trust and macro legitimacy

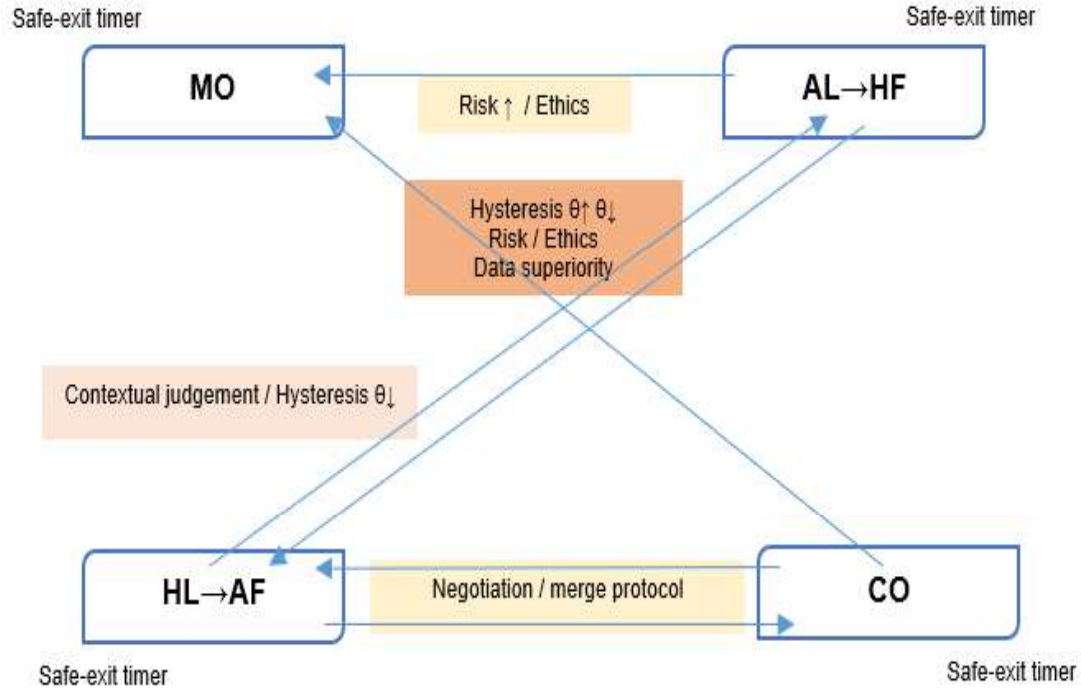
Trust studies focus on individual reliance, whereas legitimacy studies examine public accountability and regulatory compliance. These aspects can become disconnected when responsibilities are unclear. Effective auditing and governance need transparent accountability processes that connect actions to valid reasons (Raji et al., 2020). Ethics research also emphasizes that high-level principles need operational mechanisms and clear responsibilities (Mittelstadt, 2019). Recent regulations, such as the EU AI Act (Regulation (EU) 2024/1689), emphasize the need for documented human oversight in high-risk AI, underscoring the importance of clear role assignments and responsibilities. Socio-technical critiques highlight the importance of context and emphasize the need for designs that promote accountability and institutional awareness (Frimpong, 2025; Selbst et al., 2019).

For DAR, the Reversal Register and state-aligned explanations create a layer of accountability that links individual trust to overall legitimacy, making authoritative actions auditable and transparent for stakeholders and regulators.

3. The Dar Model

The Dynamic Authority Reversal Model (DAR) (Figure 1) provides a refined perspective on authority within AI-human systems, highlighting the dynamic nature of leadership roles during collaborative tasks. In contrast to conventional models that assign static roles, the DAR framework demonstrates that leadership is cyclical and adapts to current contextual cues.

Figure 2. DAR State-and-Transition Diagram (including safe-exit and hysteresis



Sources: Developed by the Author, 2025

Figure 1 shows four states (HL→AF, AL→HF, CO, MO) triggered by data superiority (AL→HF), contextual judgment (HL→AF/CO), and a risk threshold with an ethics override (MO). Timers limit dwell time, and hysteresis (θ↑/θ↓) prevents fluctuations. Transitions update a Reversal Register logging state, trigger, actors, thresholds, and explanations.

It builds on the notion of distributed leadership and examines hybrid socio-technical systems, especially in situations where participants may show limited moral reasoning or emotional intelligence. This model further investigates instances of temporary non-human leadership and the evolving dynamics of authority between humans and algorithms. Ultimately, it posits that AI can occupy roles that are subordinate, dominant, or shared in leadership responsibilities, contingent on the specific context. This underscores the notion that leadership is a dynamic exchange of influence shaped by factors such as trust, accountability, and situational variables.

3.1 States and semantics

- **HL→AF (Human Leader → AI Follower):** People make choices while AI provides suggestions, assessments, and confirmation. The responsibility rests with the individual, and the goal is to enhance their decision-making with analytical insights.
- **AL→HF (AI Leader → Human Follower):** Artificial intelligence suggests a policy or action with the ability to enforce it unless opposed; a human oversees this process and has the power to halt or reject it. Accountability locus: designated human owner; operational

lead: AI. Explanation target: operational sufficiency (confidence, alternatives, risk envelope).

- **CO (Co-Leadership):** Control is negotiated or shared; neither party holds unilateral authority. Accountability locus: shared, with explicit division of labor and a merge protocol. Explanation target: negotiation-grade rationale and conflict-resolution path.
- **MO (Mutual Override):** Either side can halt or reverse; typically bound to high-risk or ethical boundaries. Accountability locus: human (ultimate responsibility owner) with logged override causes. Explanation target: justification of override and post-hoc audit trail.

3.2 Triggers

1 Data superiority: Switch to AL→HF when the AI's predictive margin over baseline (e.g., last-best human policy) exceeds θ for k consecutive ticks or samples.
2 Contextual judgment need: Switch to HL→AF when tacit knowledge, stakeholder values, or non-codified norms dominate expected utility.
3 Risk threshold: Enter MO when tail risk/harm likelihood crosses R; require human sign-off or abort.
4 Ethics override: If policy or legal rules prevent continued autonomy, use MO or HL→AF.
Hysteresis: Transitions include stickiness to prevent oscillation—e.g., θ↑ for entering AL→HF and θ↓ for leaving, so temporary dips do not thrash the state.

3.3 Guardrails and safe-exit

- **Safe-exit timers:** Maximum dwell time in a state before a reevaluation checkpoint.
- **Override lanes:** Low-latency paths for humans to stop, reverse, or demand explanation without punitive friction.
- **Role-state logging:** The Reversal Register records state, trigger, actors, thresholds, explanations delivered, and actions taken.
- **State-contingent XAI:** Explanation surfaces built to match the active state's needs.

3.4 Metrics and telemetry

- **Reversal latency (ms or decision cycles):** Time from trigger to effective state change.
- **Hysteresis width:** Gap between entry and exit thresholds that define stickiness.

Trust-whiplash rate: Incidence of user reliance shifts (over-/under-reliance) after reversals; measured via decision adherence or correction rates.

Thrash rate: Unwanted rapid state flips per 1,000 decisions (should be near zero if hysteresis is tuned).

Phase-responsibility completion: Percentage of decisions with complete state-aligned explanations and sign-offs.

4. Propositions And Methodological Pathways

The Dynamic Authority Reversal (DAR) framework outlines measurable relationships between authority allocation and key outcomes, including effectiveness, safety, and legitimacy. This framework underscores the importance of collaboration between humans and artificial intelligence in adapting to evolving leadership conditions. Furthermore, it emphasizes the importance of essential tools that ensure these transitions are transparent and accountable.

- ✓ **Proposition 1 (Authority–Performance Fit).** Decisions made based on current information outperform those made based on static role allocations. When AI predictions significantly exceed a baseline for an extended period, the authority should transition from a human-led approach to an AI-driven one. Conversely, when human judgment, expertise, or values are more important, authority should return to Human Leaders. This approach aims to enhance key performance indicators and minimize serious errors compared to traditional human-in-the-loop approaches.

- ✓ **Proposition 2 (Safe-Exit and Acceptance).** Introducing explicit safe-exit timers improves the acceptance of AI systems by enforcing limited autonomy and predictable reviews. These timers set clear deadlines for when human oversight will take over, reducing overreliance on AI while maintaining efficient processes. This method enhances decision-making and reduces incidents without compromising operational efficiency.
- ✓ **Proposition 3 (Hysteresis and Stability)** Establishing asymmetric entry and exit thresholds, commonly referred to as hysteresis, is an effective strategy for minimizing undesired state fluctuations while maintaining optimal performance. A wider hysteresis lowers churn and thrash rates, but if the bands are too broad, it may delay needed reversals. The approach aims for a solution in which stability improvements outweigh performance losses to some extent.
- ✓ **Proposition 4 (State-Contingent Explainability and Justification).** Tailored explanations for active leadership enhance the completeness and auditability of the decision rationale. By ensuring operational sufficiency in AL→HF (confidence, alternative actions, risk management), providing diagnostic support in HL→AF, offering negotiation-quality rationale in CO, and justifying overrides in MO, we improve phase responsibility completion and increase auditor satisfaction compared to generic explanation methods.
- ✓ **Proposition 5 (Reversal Latency and Safety).** Reducing the time between a valid trigger and handover minimizes risk in sensitive workflows. Tools that shorten detection-to-handover intervals should lead to fewer near-miss events and lower tail-risk events, while accounting for task difficulty and case mix.
- ✓ **Proposition 6 (Transitional Accountability and Legitimacy).** Maintaining a Reversal Register that records states, triggers, thresholds, justifications, and human endorsements for every decision can enhance stakeholder trust. Companies that keep this register typically obtain better audit ratings, experience fewer reversals in review outcomes, and have stronger legal protections compared to those that do not maintain such a register.

To evaluate these propositions efficiently, systems should provide a minimal telemetry set that includes the current state, trigger type, relevant thresholds for entry and exit, dwell time, safe-exit status, explanation class delivered, and the sequence of actor actions (approval, override, demand for elaboration). From this data, researchers can calculate metrics such as reversal latency, hysteresis width, trust-whiplash, thrash rate, and phase-responsibility completion, and benchmark results against a static HITL configuration. This framework facilitates the accumulation of comprehensive evidence while supporting the analysis of diverse experimental designs. These include randomized threshold bands, A/B tests using simulators, and field pilots.

5. Implications And The Authority-State Playbook

To ensure effective leadership in socio-technical systems, it is essential to integrate clarity from the outset rather than adding it later as a control measure. The Authority-State Playbook (ASP) implements this by linking policies to telemetry and defining leadership roles, authorization processes, and justification methods beforehand. The ASP starts with a role matrix that assigns specific human owners to areas of authority and followership, reducing ambiguity around oversight. It also outlines triggers—such as data thresholds, risk boundaries, and ethics rules—and includes settings for entry and exit points to create a stable decision-making environment. Temporal boundaries are crucial as well. Safe-exit timers prevent prolonged durations in any state without requiring assessment, ensuring human oversight. The Reversal Register monitors changes in state, triggers, and justifications, facilitating accountability and ongoing enhancement based on previous experiences.

For individuals in leadership roles, the explanations must correspond with the current situation. In cases like AL→HF or HL→AF, understanding model confidence, risk, and rationales is crucial for making informed decisions. Introducing clear leadership within AI frameworks effectively mitigates dependency issues and boosts team morale by allowing for predictable transitions. Adjusting parameters such as hysteresis bands and timers is crucial for ensuring system stability without compromising responsiveness. Providing customized explanations that meet specific needs can help

manage cognitive load and encourage informed decision-making. Additionally, employing the Reversal Register as a governance tool not only fosters accountability but also nurtures a culture of learning and risk awareness.

At the policy level, the ASP adheres to existing regulations concerning human oversight and documentation for high-risk AI applications. It underscores the importance of transitional accountability, enabling regulators to assess whether justifications align with state requirements and to comprehend the decision-making processes involved. This method facilitates bounded autonomy while upholding human responsibility, acknowledging that reversible AI leadership can serve as an effective strategy for achieving this balance. Furthermore, selecting suitable triggers and settings strikes a balance between innovation and safety, avoiding a one-size-fits-all regulatory framework.

6.Conclusions

Organizations need not just improved models but also better handover processes. The Dynamic Authority Reversal (DAR) framework reconceptualizes human–AI collaboration as a challenge of temporal coordination, offering the necessary transition logic to clarify leadership roles, timing, and reasons. By delineating four leadership states—Human Leader to AI Follower, AI Leader to Human Follower, Co-Leadership, and Mutual Override—alongside the principled triggers that enable transitions between them—such as data superiority, contextual judgment, risk thresholds, and ethical considerations—DAR effectively links performance with accountability.

To reduce the risks of oscillation and drift, safeguards such as safe-exit timers and hysteresis are used. Additionally, state-contingent explainability ensures that appropriate justifications are provided during critical moments. Significantly, the Reversal Register ensures that authority changes are auditable by associating each decision with its respective state, trigger, thresholds, and approvals. This connection promotes micro-level trust calibration with macro-level legitimacy.

The straightforwardness of DAR's tools promotes practical implementation and evaluation, offering a concrete and verifiable approach to moving from generic "human oversight" to an auditable leadership structure. For researchers, it offers testable propositions regarding the impacts of reversal latency, stability, and legitimacy. For practitioners and regulators, it serves as a governance-ready framework that preserves ultimate human accountability while enabling reversible AI leadership in areas where it is most advantageous.

References

Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for Human–AI Interaction. In CHI '19. https://doi.org/10.1145/3290605.3300233

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.14779

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. https://doi.org/10.1145/3449287

Chen, W., & Zhang, J.-H. (2022). Does Shared Leadership Always Work? a state-of-the-art Review and Future Prospects. Journal of Work-Applied Management, 15(1), 51–66. https://doi.org/10.1108/jwam-09-2022-0063

Frimpong, V. (2025). Algorithmic authority and the complexities of delegated decision-making: Case studies on ethical challenges for 21st-century leadership. International Journal of Organizational Leadership, 14(3), 637–655. https://doi.org/10.33844/ijol.2025.60525

European Union. (2024). Artificial Intelligence Act. Regulation (EU) 2024/1689, 13 June 2024. Official Journal, 12 July 2024.

Liao, Q. V., Gruen, D., & Miller, S. (2020a). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376590

Liao, Q.V., Gruen, D., & Miller, S. (2020b). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Maritsa, E., Goula, A., Psychogios, A., & Pierrakos, G. (2022). Leadership development: Exploring relational leadership implications in healthcare organizations. International Journal of Environmental Research and Public Health, 19(23), 1–14. https://doi.org/10.3390/ijerph192315971

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. (2019). Principles alone cannot guarantee the ethical development of AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How Different Explanation Classes Impact Trust Calibration: The Case of Clinical Decision Support Systems. International Journal of Human-Computer Studies, 169, 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

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. https://doi.org/10.1145/3411764.3445315

Rahwan, I. (2019). Machine behaviour. Nature, 568(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

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. https://doi.org/10.1145/3351095.3372873

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 - FAT* '19, 59–68. https://doi.org/10.1145/3287560.3287598

Bollaert, M., Augereau, O., & Coppin, G. (n.d.). Measuring and Calibrating Trust in Artificial Intelligence Measuring and Calibrating Trust in Artificial Intelligence. Retrieved August 28, 2025, from https://hal.science/hal-04493669v1/file/Trust_Calibration_in_Artificial_Intelligence-1.pdf

Wu, Q., Cormican, K., & Chen, G. (2020). A meta-analysis of shared leadership: Antecedents, consequences, and moderators. Journal of Leadership & Organizational Studies, 27(1), 49–64. https://doi.org/10.1177/1548051818820862