THE OVERRIDE HIERARCHY: HUMAN AUTHORITY PATTERNS FOR SAFETY-CRITICAL AI

***Hrozian Y.**, Founding Designer, Lumos AI, San Francisco, California, USA*
***Chebotarova I.,** Senior Lecturer of the MST Department, KhNURE*

Abstract. *This paper introduces the Override Hierarchy: a four-tier framework (Defer, Confirm, Reject, Reverse) for designing human authority over AI in safety-critical systems. Drawing on FDA 2025 guidance and human-AI interaction research, it argues that not all overrides are equal: interface design must encode authority gradients explicitly. The framework gives UX teams a vocabulary for override flows in clinical AI products.*

Keywords: *Override Hierarchy, human-AI interaction, clinical AI, safety-critical AI, human oversight, authority patterns.*

The rise of AI in safety-critical settings has made human override a regulatory and clinical imperative, but rarely a design priority [1-4]. The FDA's January 2025 draft guidance on AI-enabled device software [5] and the EU AI Act both require human oversight, yet the design vocabulary remains scattered. Many AI products treat override as a binary, when clinicians, operators, and regulators need graded authority patterns. The Override Hierarchy distinguishes four tiers, each with different interface requirements, audit consequences, and downstream signal (table 1).

Table 1 – The Override Hierarchy

Tier	What the user is doing	Example interface
Defer	Accepting AI output but signaling reduced confidence; flagging case for review.	"Accept and route for review" action; case enters a peer-review queue.
Confirm	Explicitly endorsing AI output and taking authorial responsibility.	Two-stage confirmation; signed timestamp; audit log entry.
Reject	Declining AI output without correcting it; case proceeds without AI input.	One-click rejection with optional reason; output marked as rejected.
Reverse	Actively contradicting AI output and substituting a different decision.	Counter-decision UI with rationale capture; surfaces to feedback loop.

Defer is the most under-implemented tier. The clinician neither endorses nor rejects the AI's output but routes it for additional review: "I'm uncertain, escalate." [6] confidence-calibration framework formalizes the pattern. Low-confidence outputs trigger secondary review, and high-confidence predictions are overridden at only 1.7%. The closest real-world parallel is independent double reading in radiology. Absent a Defer tier, products force false confidence by giving clinicians no middle ground.

Confirm carries weight beyond endorsement. It transfers authorial responsibility from model to human. Ambient AI scribe products like Abridge and Nuance DAX [7] require explicit clinician sign-off on each draft note; the timestamp and audit log entry anchor liability under the FDA's lifecycle-management framework. Two-stage confirmation, where the clinician sees what was edited and explicitly attests,

is the load-bearing pattern. Without a clear attestation surface, the chain of authorial responsibility breaks regardless of model output quality.

Reject is a clean refusal. The AI's output does not flow into the clinical record. Reject differs from Defer because no further review is requested; it differs from Reverse because no alternative is substituted. Epic's Best Practice Advisories illustrate the alert-fatigue cost of rationale-on-every-reject: clinicians dismiss in bulk without engaging with the model's reasoning. The right default is rationale-optional with structured-reason categories ("model wrong," "context inappropriate," "false alarm"), so the platform can distinguish model error from contextual change.

Reverse is the most information-rich tier. The clinician actively contradicts the AI and substitutes a different decision. This is the teaching signal the model can learn from. Pathology AI platforms such as PathAI and Paige illustrate continuous-learning workflows where pathologist feedback flows back into ongoing model refinement. Amershi et al.'s [8] error-and-recovery guidelines (G7-G11) frame override around recovery; the Override Hierarchy names the authority gradient recovery alone cannot capture.

Translating the hierarchy into a specific product requires three choices: which tiers the system supports, what friction each tier carries, and where the signal flows back. Different surfaces need different subsets: what works for ambient scribes will not work for sepsis alerts. Adding tiers adds complexity, and busy clinicians prefer fast paths. The hierarchy is a vocabulary, not a checklist.

Future work should examine how the FDA's Total Product Lifecycle and the EU AI Act risk classifications map to the hierarchy. Cross-organization studies of override telemetry would illuminate which patterns sustain trust over months of deployment [9]. For product teams designing AI in safety-critical contexts, naming the tiers makes them designable, auditable, and discussable: the foundation of trustworthy AI deployments.

References

1. Borovynska, Y., & Vovk, O. (2024). Investigating the vision of AI driven website builder in user interface components. Jóvenes en la ciencia, (26). https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4233/3714.

2. Chebotarova, I., & Silchenko, V. (2024). Intelligent text recognition when creating audio books for blind people. Jóvenes en la ciencia, (26). https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4232/3713.

3. Kaluhin, N., Vovk, O., & Chebotarova, I. (2024). The impact of artificial intelligence on future of humanity. Jóvenes en la ciencia, (26). https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4235/3716.

4. Khlynyna, S., Vovk, O., & Chebotarova, I. (2024). Prospects for using artificial intelligence for book layout. Jóvenes en la ciencia, (26). https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/4236/3717.

5. U.S. Food and Drug Administration. (2025). Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (Draft Guidance). fda.gov/regulatory-information/search-fda-guidance-documents.

6. Yu, Y., Gomez-Cabello, C.A., Haider, S.A., et al. (2025). Enhancing Clinician Trust in AI Diagnostics: A Dynamic Framework for Confidence Calibration and Transparency. Diagnostics, 15(17), 2204. doi.org/10.3390/diagnostics15172204.

7. Shah, K.P., & Johnson, K.B. (2025). The Ambient AI Scribe Revolution – Early Gains and Open Questions. JAMA Network Open, 8(10). jamanetwork.com/journals/jamanetworkopen/fullarticle/2839544.

8. Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for Human-AI Interaction. 2019 CHI Conference on Human Factors in Computing Systems. (p. 1-13). ACM. doi.org/10.1145/3290605.3300233.

9. Tun, H.M., Rahman, H.A., Naing, L., & Malik, O.A. (2025). Trust in Artificial Intelligence-Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review. Journal of Medical Internet Research, 27, e69678. jmir.org/2025/1/e69678.