WHEN THE ALGORITHM STOPS LISTENING

Selective Ground Truth in Adaptive Medical AI: Preventing Expert Drift in PCCP-Governed Systems

Mark Paull · Independent Researcher · Montreal, Canada · 60th year with type 1 diabetes · 100,000+ insulin injections · markpaull56@gmail.com

STUDY TYPE: Conceptual framework analysis

THE CORE PROBLEM: Adaptive systems may mistake expert correction for non-adherence — then retrain on that mistake.

BACKGROUND & PROBLEM

Adaptive clinical AI, including automated insulin delivery, increasingly retrains on real-world behavioural data. A valid override may reflect context the model cannot see. Coded as non-adherence, expertise becomes error.

METHODS / FRAMEWORK ANALYSIS

Conceptual systems analysis of adaptive AID retraining pipelines under the U.S. FDA Predetermined Change Control Plan (PCCP). Override events were treated as information-bearing clinical interventions, not behavioural deviations.

Residual Governance Effort (RGE)
The human correction required to keep an adaptive system safe when clinical context sits outside the model's feature space.

Proposed operationalization: governance-informed overrides requiring human relabeling per retraining cycle (not yet computed).

SELECTED REFERENCES

Sambasivan et al. Data cascades in AI. CHI 2021.
Sculley et al. Hidden technical debt. NeurIPS 2015.
U.S. FDA. PCCP guidance, AI device software. 2024.

THE FAILURE FORK

The same override. The label decides everything.

A VALID OVERRIDE
"I skip this correction — I'm about to exercise."

how is it labeled?

WRONG PATH	RIGHT PATH
LOGGED AS NON-ADHERENCE	TAGGED WITH REASON CODE
ENTERS RETRAINING AS "ERROR"	GOVERNANCE-INFORMED RETRAINING
MODEL DOWN-WEIGHTS CORRECT BEHAVIOUR	EXPERT SIGNAL PRESERVED
EXPERT DRIFT	MODEL STAYS ALIGNED

MY OWN OVERRIDE

I override an insulin recommendation because I know I am about to exercise. The system logs "non-adherence." I was not non-adherent — I was right. Retrained on that label, the model learns to distrust the judgment that kept me safe.

EXPERT DRIFT
A retraining failure in which valid human correction is misclassified as user error, degrading ground truth over time.

CONCEPTUAL ANALYSIS

Override events encode high-resolution clinical judgment unavailable to the model's feature space. Reduced to non-adherence, retraining optimizes toward behavioural conformity, not outcome safety.

KEY FINDING: Without context-aware override labels, adaptive AI may learn against the expertise keeping patients safe.

WHAT OVERRIDES MAY CONTAIN

- Sensor instability or signal artifact
- Illness, stress, exercise, competing signals
- Situational constraints the model cannot see
- Judgment built over years of self-management

REQUIRED SAFEGUARDS

Developers	Structured override reason coding.
Regulators	Override-labeling in PCCP filings.
Clinicians & Patients	Log overrides as expertise.