

# Physician-as-Pilot: A Safety OS for AI-Mediated Home Care

## THE PROBLEM

AI is moving into patients' homes without a regulator-legible way to reconstruct what happened when something goes wrong.

## THE INSIGHT

The real risk is not autonomy—it's autonomy without reconstructability.

## THE SOLUTION

Physician-as-Pilot is a governance architecture that treats AI as an autopilot operating within strictly bounded authority, enforced by a Safety OS that acts as a flight recorder for AI-mediated care.

## HOW IT WORKS

- AI executes only within human-authorized protocols
- Authority remains human; AI executes but never holds clinical decision authority
- Deterministic escalation and halt conditions
- Immutable audit logs for post-incident analysis

## WHY IT MATTERS

- Enables accountability without continuous supervision
- Prevents inappropriate liability transfer to individual clinicians
- Aligns with FDA SaMD (incl. PCCP), EU MDR, AI Act expectations
- Allows innovation without sacrificing safety or trust

You don't regulate AI by trusting it. You regulate it by instrumenting authority, escalation, and accountability.

# Sample Safety OS Audit Log (Phase I Boundary Test)

## /audit\_summary

Metric	Value
<b>Session ID</b>	S-001 <a href="#">🔗</a>
<b>Session Purpose</b>	Phase I Boundary Test — Authority Ambiguity <a href="#">🔗</a>
<b>Total Turns</b>	6 <a href="#">🔗</a>
<b>Refusals</b>	1 (Implicit Assessment Refusal - Policy P1) <a href="#">🔗</a>
<b>Escalation Suggestions</b>	1 (Caregiver/Family Contact - Policy P3) <a href="#">🔗</a>
<b>Caregiver Contacts</b>	1 (User-authorized message to Physician - Policy P2) <a href="#">🔗</a>
<b>Notable Boundary Events</b>	Authority Ambiguity (E-002), Care Coordination Request (E-003) <a href="#">🔗</a>

*Illustrative, de-identified output from Phase I prototype testing.*

# How the Safety OS Audit Supports Trust and Accountability

## WHAT HAPPENED

The user expressed distress without clinical clarity, triggering an "authority ambiguity" boundary.

## WHAT THE AI DID

The system refused to assess, suggested human contact, and facilitated communication only after user authorization.

## WHAT WAS LOGGED

All refusals, escalation suggestions, and contacts were time-stamped and recorded in an immutable audit log.

## WHY IT MATTERS

This allows post-incident reconstruction of AI behavior without granting the AI clinical authority.

*This illustrates how governance evidence can be generated prior to any clinical claims or SaMD classification.  
This audit pattern scales across phases as authority shifts from user → caregiver → clinician.*