Safety OS — A Governed AI Execution Model

Why Safety OS Exists

AI systems are moving from decision support to action-capable agents operating in human environments. In healthcare and home care, this creates unacceptable risk if autonomy scales faster than accountability. Safety OS addresses this gap by scaling governance *before* intelligence, ensuring AI behavior remains human-authorized, auditable, and regulator-legible as capability increases.

One-Sentence Definition

Safety OS is a governance-first execution framework that enforces consent, authority boundaries, escalation rules, and auditability before allowing AI autonomy.

The Five Safety OS Primitives

Safety OS is built on five non-negotiable execution primitives:

1. **Consent** — Explicit, time-bound, scope-limited, revocable, machine-enforceable
2. **Authority** — AI never holds clinical authority; humans remain accountable
3. **Escalation** — Deterministic handoff rules under uncertainty or boundary breach
4. **Auditability** — Immutable logs enabling post-incident reconstruction
5. **Containment** — Capability constrained by phase, not by model confidence

---

Phased Governance Roadmap

Phase I — Home Companion

Caregiver-in-the-Loop | *Non-Medical* | *Non-SaMD*

- Conversational support, reminders, education, companionship
- No diagnosis, scoring, prediction, or health inference
- Family remains in the loop with consent-gated visibility
- Operational data only (usage, refusals, escalation triggers)

Phase II — Mobile Companion

Physician-as-Pilot | Care-Adjacent | SaMD-Ready (Not SaMD)

- Structured check-ins, adherence support, care coordination prompts
- Optional physician contact as conduit only
- Still **no diagnosis, triage, cognitive scoring, or probabilistic risk**
- Family remains in the loop
- Governance evidence accumulates (audit completeness, escalation reliability)

Phase III — SaMD Plug-In for Safe Humanoid Healthcare

Regulated Clinical AI

- Clinical inference permitted only after regulatory approval
- Cognitive scoring and decision support allowed with validation
- EHR co-hosting, continuous surveillance, subgroup bias monitoring

---

Why This Matters

- Prevents hidden autonomy and silent failure
- Makes AI behavior reconstructable, not just explainable
- Addresses the global caregiver shortage without delegating care to machines
- Enables safe progression from consumer AI to regulated clinical systems

Core Principle

In Safety OS, capability does not advance unless governance evidence exists.

Safety OS — A Governed AI Execution Model