Safety OS — A Governed AI Execution Model

1. The Problem Safety OS Solves

Most AI frameworks optimize model performance first and attempt governance later. In healthcare, this inversion creates ethical risk, regulatory fragility, and loss of trust. As AI agents move into homes, caregiving, and clinical environments, the primary failure mode is not inaccuracy — it is **unaccountable action**.

Safety OS reframes AI as an execution problem, not a modeling problem.

---

2. From Design Thinking to Governed AI Execution

Traditional innovation models (e.g., IDEO's Inspiration → Ideation → Implementation → Iteration) assume humans are the primary decision-makers. AI agents break this assumption by acting continuously, autonomously, and at scale.

Safety OS extends design thinking into **governed execution**, adding:

- Continuous enforcement (not static guidelines)
- Machine-readable constraints
- Evidence-based progression gates
- Post-incident reconstructability

---

3. The Safety OS Execution Stack

Safety OS operates above models and below interfaces:

- **Policy Layer** — consent rules, authority scopes, escalation conditions
- **Control Layer** — deterministic enforcement of boundaries
- **Execution Layer** — voice agents, mobile companions, humanoids
- **Evidence Layer** — immutable logs, incident traces, compliance artifacts

Models may change. Governance does not.

---

4. The Five Governance Primitives (Expanded)

Consent

- Explicit opt-in, revocable, purpose-bound
- Separate consent for patient, caregiver, physician
- Machine-enforced, not policy-only

Authority

- AI never diagnoses, prescribes, or prioritizes independently
- Authority explicitly assigned to caregiver or clinician
- No implicit delegation through “recommendation creep”

Escalation

- Deterministic triggers (uncertainty, refusal, ambiguity)
- Human handoff required, logged, and traceable

Auditability

- Every action reconstructable after the fact
- Supports regulators, clinicians, families, and courts

Containment

- AI capability bounded by phase, not confidence
- No silent expansion of function

---

5. Phased Governance Roadmap (Detailed)

Phase I — Home Companion

- Voice-based companionship and education
- Family-in-the-loop with consent and access control
- No health inference or scoring
- Generates real-world governance data without PHI

Phase II — Mobile Companion (Physician-as-Pilot)

- Adds mobility and structured care-adjacent interactions
- Physician interaction is optional and human-led
- Still not a medical device
- Governance evidence required before progression

Phase III — SaMD Plug-In

- Formal regulatory submission
- Clinical inference allowed under protocol
- EHR cloud co-hosting (e.g., Epic)
- Bias audits, drift monitoring, and recertification required

---

6. Generalizability as a Governance Obligation

Safety OS rejects the idea that models are "universally generalizable."
Instead:

- Performance must be validated locally
- Deployment contexts must be declared
- Drift triggers recalibration or rollback

Generalizability is governed, not assumed.

---

7. Why Safety OS Is Different

- It treats AI as a socio-technical system
- It aligns with emerging regulator expectations
- It scales accountability before autonomy
- It directly addresses caregiver scarcity without automating care

---

8. Final Principle

If authority cannot be traced, capability must not advance.
Safety OS operationalizes this principle across consumer, care-adjacent, and clinical AI.