FIELD GUIDE · THE 2026 HINTON LECTURES · CLINICAL TRANSLATION

# Clinical AI Safety Checklist

Twelve checks before an AI tool touches a patient, plus ten questions for every vendor.

## FIVE SAFETY CHECKS (adapted from Paul Christiano, Night 3)

| Specification | Alignment | Capabilities | Control | Security |
|---|---|---|---|---|
| Written intended use, populations, exclusions | Validated on our own patients | What else it can do; what each update changes | Who watches, who overrides, how to stop safely | Who can reach data and model; revoke fast |

## CHECKLIST

- ☐ Intended use is written, including who it is not for
- ☐ Local validation done on our own population
- ☐ Version is locked; updates trigger review
- ☐ A named clinical owner signed before go-live
- ☐ Staff trained on limits and automation bias
- ☐ Patients told when AI meaningfully shapes care
- ☐ Outcomes monitored, including by subgroup
- ☐ AI events reportable in our safety system
- ☐ Downtime plan exists and has been drilled
- ☐ Data access mapped and revocable within an hour
- ☐ Someone can pause it, in writing, without a meeting
- ☐ Independent review at least yearly

## TEN QUESTIONS FOR ANY AI VENDOR

01 What was the model optimized for?

02 Which populations was it validated on?

03 How and when do you notify us of changes?

04 Can we lock the version we validated?

05 What does it do when it is uncertain?

06 What data leaves our walls, and where?

07 How do we pause or remove it in an emergency?

08 Who reviews your safety testing besides you?

09 What incidents have you had, and what changed?

10 Who signs for its safety on your side?

**The one question behind all of it**

"This tool was optimized for ____, and the outcome I care about is ____." If the blanks differ, monitor the gap.

Every clinical AI tool needs a named owner with written authority to pause it.

Independent educational resource by Harvey Castro, MD. Clinical application is the author's own analysis, informed by the 2026 Hinton Lectures (The AI Safety Foundation; lecturer Paul Christiano) and by Joint Commission/CHAI, FDA, NIST and WHO guidance. Not affiliated with or endorsed by The AI Safety Foundation. Hinton Lectures is a trademark of The AI Safety Foundation. Educational only; not medical or legal advice. Reviewed November 27, 2026.

**Full guide: https://www.harveycastromd.com/blog-posts/hinton-lectures-2026-field-guide**