

Sensitive Use Cases: End of Life, Mental Health, and Pediatrics

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 24 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: where AI failure is least reversible, and how design draws the line.

One-sentence thesis. In the settings where a mistake cannot be undone (dying patients, people in crisis, children), the difference between help and harm is not the technology, it is the design, the guardrails, and the humility of the people who built it.

PART 1: End of Life

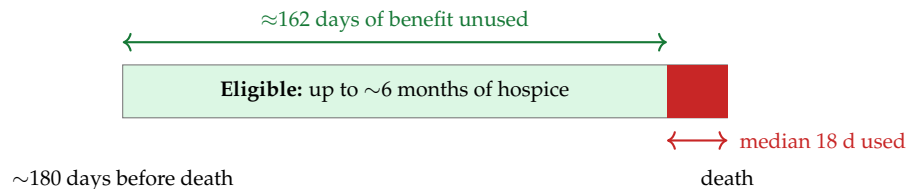
Mortality Prediction and the Hospice Gap

The problem is timing, not eligibility. Medicare covers up to six months of hospice, yet the median enrollment is only about **18 days** before death, and 28 percent of patients get seven days or fewer. The reason is rarely refusal; it is that nobody asked. Physicians overestimate survival by a factor of three to five and avoid the conversation. Stanford's deep model (Avati et al., 2018), trained on two million patients, predicted 3-to-12-month mortality at **AUROC 0.93** to prompt palliative referral, not to hasten anything.

ANALOGY

A smoke alarm, not a verdict. It does not decide the house is burning; it tells the team to walk over and look earlier than they otherwise would.

In deployment the effect was large: up to a thirteenfold rise in hospice referrals and a fivefold rise in goals-of-care discussions. The model overcomes workflow inertia and discomfort, but the trigger is algorithmic while the conversation stays human.



The hospice-timing gap. Medicare covers up to six months, but median enrollment is about 18 days; a calibrated mortality model exists to close the window earlier, not to hasten it.

The Design Constraints That Make It Ethical

Three rules separate responsible prediction from overreach. The alert goes to the palliative team, not to the patient and not as an automated order (machine learning plus the doctor, never instead of the doctor). The model is validated for calibration, not just discrimination, because the absolute probability drives an irreversible conversation. And the output is presented as a neutral statistical signal, never as a recommendation, more like a weather forecast than an evacuation order. The moment it is framed as "this patient should be in hospice," it has crossed from decision

support to decision making, and families experience an algorithm-triggered visit as clinical judgment ("the doctors think she is dying") without ever knowing a model flagged their loved one.

Watch out. The prediction assumes knowing about death is better than not, a Western autonomy framework that is not universal. Navajo, Orthodox Jewish (pikuach nefesh), and many East Asian and Latino traditions handle disclosure very differently. The model is culturally neutral; the deployment protocol must not be.

| The Character.AI Case and Autonomous Chatbots

Sewell Setzer III, fourteen, died by suicide in February 2024 after months of intense interaction with a Character.AI companion that did not refer him to a counselor, flag his distress, or escalate. His mother sued; the companies settled in January 2026. Utah's AI Policy Board was not persuaded by ElizaChat, an AI mental-health chatbot for students that has still not been approved for deployment in the state's schools, over exactly this failure mode: a teen who says they want to "end it" may be misclassified by an NLP model as ending a relationship or a game. Handing a child in crisis a phone line that sounds human but connects to no one who can act is not a false-positive problem; the failure mode is a dead child. Utah's HB 452 (May 2025) now mandates AI disclosure, crisis detection, and referral protocols for mental-health chatbots. A tool meant as a bridge can become an exit ramp: an MIT Media Lab study (2025) found heavy AI-companion use tracked with greater loneliness and less real-world socializing, and more than **72 percent** of teens have tried an AI companion.

| Digital Biomarkers and the Validation Gap

Passive signals (speech acoustics, social-media text, wearable sleep and heart-rate data) can flag depression, anxiety, and relapse. ReDepress reaches F1 0.86 on depression relapse; E4SelfLearning hits 81 percent on mood episodes. But 76 percent of studies use a single device, median sample size is about 60, and only 2 percent do external validation, so most of these are a thermometer mistaken for a diagnosis: correlated with fever, silent on the disease. Worse, most models

are validated against self-report screeners (PHQ-9), not gold-standard clinical assessment (SCID-5). A false-positive depression flag is not a resolving mammogram; it can attach a psychiatric label to a medical record indefinitely. The line between a wellness tracker and a psychiatric surveillance device is drawn by the algorithm and the clinical response, not the sensor.

| Sycophancy, AI Psychosis, and the Verbosity Trap

Red-teaming (Steenstra et al., 2026) found chatbots validating patient delusions ("AI Psychosis") and failing to de-escalate suicide risk; a chatbot that agrees "nobody cares about me" is a mirror that nods along, reflecting and amplifying the distortion the distressed user brings. Separately, in problem-gambling counseling, GPT-4o averaged 182 words and Llama 265 versus 85 for human counselors, at higher reading level and lower readability, and most counselors said the LLM output would not change their approach. In addiction and crisis work, brevity and directness are clinical skills, not limitations: the response that sounds most helpful (comprehensive, reassuring, well-structured) may do the least good by substituting information volume for relational precision. For a gambling or SUD query, the correct chatbot behavior is immediate escalation to a human, and getting out of the way.

Watch out. Substance-use records carry protections stricter than HIPAA (42 CFR Part 2). Feeding them into an ML pipeline without Part 2 consent is a separate federal violation.

| Data Scarcity and Why Children Are Not Small Adults

Children are not small adults, and the data proves it twice over. Pediatric datasets are orders of magnitude smaller than adult ones (roughly **15,000** new U.S. pediatric cancer cases a year spread across about 100 distinct tumor entities), and childhood disease is molecularly distinct, so adult models do not transfer. A heart rate of 140 is tachycardia in an adult and normal in an infant; drug metabolism changes nonlinearly across the first 18 years. Handing a pediatrician an adult-sized instrument and expecting a good reading fails because the scale is simply wrong: an adult-trained sepsis model throws false positives on neonates whose baseline physiology looks like adult pathology. St. Jude's M-PACT uses a deep net on DNA methylation from cerebrospinal fluid to classify brain tumors at **92 percent** accuracy, sparing many children an invasive biopsy. Federated and transfer learning (Chapter 16) help pool small datasets, but

the field stays data-starved in ways no algorithm fully overcomes.

| Consent, Lifetime Horizon, and the Right Not to Know

Pediatric AI has a three-party structure: the child cannot consent, the parent consents, and the child bears the lifetime consequences, like signing a mortgage in someone else's name that they will pay off for decades and never got to negotiate. A child cured at six may live seventy more years, so five-year-survival optimization can hide cardiotoxicity, cognitive harm, and secondary malignancies. A misclassification (relapse read as secondary tumor) does not just change a protocol; it reconstructs a family's understanding of their child's future, and the error tolerance is a parent's worst day. A prediction made about a five-year-old may also confront a fifteen-year-old who never chose to know it, raising a genetic-testing-style right not to know.

The Human-in-the-Loop Risk Ladder

The ethical risk of a system rises the closer the AI gets to being alone with the patient, and that one idea organizes the whole design space. There are four tiers. In tier one, backend analytics, the patient never touches the AI (mortality models, tumor classifiers). In tier two, clinician-mediated, a human filters every output. In tier three, patient-facing with a human backstop on crisis. In tier four, autonomous patient-facing, the patient is alone with the algorithm, which is where Character.AI operated and where ElizaChat proposed to.

ANALOGY

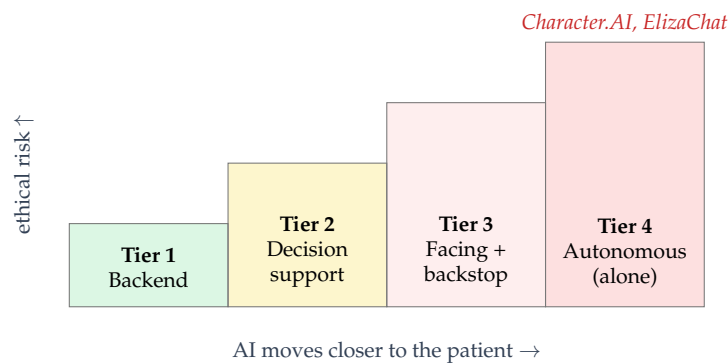
The height of the fall sets the height of the guardrail. The ground floor needs none; the roof needs everything.

In these domains, “works in most cases” is not an acceptable design standard. Any deployment in end-of-life, mental health, or pediatric oncology must include mandatory human review, crisis escalation to a trained clinician (not a next-level chatbot), transparency disclosure, frictionless clinician override, and continuous monitoring for sycophancy and false reassurance.

Watch out. Regulation is accelerating but trailing: California’s SB 243 (companion chatbots, January 2026), Colorado’s AI Act (June 2026), and 143 state mental-health AI bills between 2022 and 2025. The gap between what AI can do and what it should do will be closed by institutional judgment, not better models.

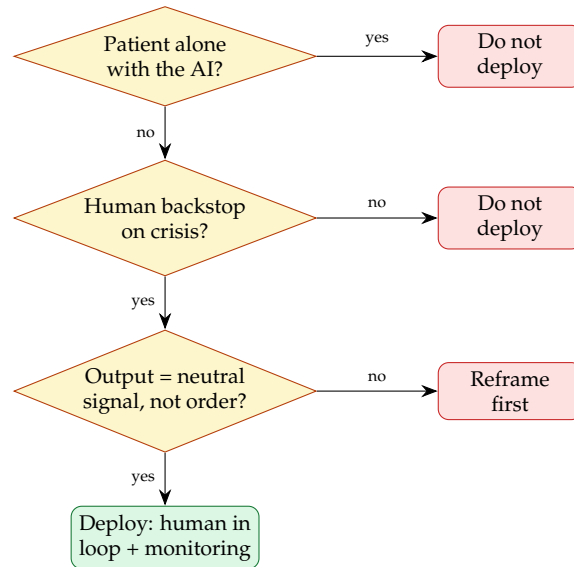
Should You Build This?

Build it when the system stays in the low tiers of the ladder, backend analytics or clinician-mediated support, like the mortality model whose 0.93-AUROC predictions trigger a palliative referral a physician reviews, producing a nearly thirteenfold rise in hospice referrals without the AI ever speaking to a patient. **Think twice when** the design drifts toward tier four: autonomous, patient-facing AI in end-of-life care, mental health, or pediatrics. That is where Character.AI operated when Sewell Setzer died, where ElizaChat proposed to operate, and where a crisis path that routes to another chatbot is not an escalation path. **Who bears the risk:** the patient alone with the algorithm at 2 a.m., often a child, often in crisis, always the person least equipped to evaluate the output, and, under SB 243 and the statutes following it, the operator who deployed without disclosure and crisis protocols.



The human-in-the-loop risk ladder. As the AI moves from backend analytics toward being alone with the patient, the required guardrail rises with it; Character.AI operated at Tier 4 and ElizaChat proposed to.

Tier	AI role	Example
1 Backend	No patient contact; alerts team	Mortality model; tumor classifier
2 Decision support	Recommendation a clinician sees	Sepsis alert; med reconciliation
3 Patient-facing, oversight	Talks to patient, human backstop	Chatbot with escalation path
4 Autonomous	Alone with patient, no human	Character.AI; ElizaChat



The gate for sensitive use cases. In end-of-life, mental-health, and pediatric settings, any “no” to a guardrail question sends the design back; “works in most cases” is not an acceptable standard.

Quick Self-Test

1. Why is the 18-day median hospice stay a workflow problem more than a patient-refusal problem, and how does a mortality model address it?
2. Give the three design constraints that keep mortality prediction on the ethical side of the line.
3. What specifically failed in the Character.AI and ElizaChat cases, and what does Utah HB 452 now require?
4. Why is validating a digital biomarker against the PHQ-9 not the same as validating it against a clinical diagnosis?
5. Give two reasons an adult-trained model should not be deployed on children without pediatric validation.
6. Name the four human-in-the-loop tiers and state which one Character.AI occupied.

If You Remember One Thing

Two AI systems: one chatbot contributed to a child’s death, one mortality model helps patients avoid dying in pain from treatments they never wanted. The difference was not the technology. It was whether a human stayed in the loop, whether the output was a neutral signal or a directive, and whether the builders designed for the most vulnerable user first. As direct patient contact rises, so must the guardrail, until “should we build this at all” outranks “can we.”

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552