AI Ethics

Structured Controversy

Macy Foundation / Millennium Conference • Preclinical M1/M2 • November 2026

“Should our hospital deploy a no-opt-out AI sepsis prediction system?”

30 minutes • pairs • no technology required • closes with a scored four-principle debrief

The format: students argue an assigned side, then switch and argue the opposite side, then design a third way together. The goal is not consensus — it is the skill of engaging a real ethical disagreement without collapsing into one side.

The Scenario

Riverbend Health, a five-hospital system, is deciding whether to activate “SepsisGuard,” a machine-learning early-warning system, across all admitted adult patients. Note: All numerical values are hypothetical scenario parameters created for educational purposes.

What the tool does

1. Every hour, the model analyzes each admitted patient's vital signs, labs, and nursing documentation, and produces a sepsis risk score from 0–100 — a probabilistic prediction of sepsis risk over the next 12 hours.
2. When a score crosses the threshold, the bedside nurse and covering clinician are paged. Protocol requires bedside evaluation within 30 minutes — and, if concern is confirmed, lactate, blood cultures, and consideration of early antibiotics and fluids.

What happens to patients downstream

3. **True positives.** Patients with early sepsis are evaluated and treated hours sooner. In external studies of comparable systems, earlier treatment is associated with meaningfully lower sepsis mortality when clinicians respond promptly.
4. **False positives.** At the planned threshold, roughly three of every four alerts will not be sepsis (PPV ≈ 25%). Those patients get an unplanned evaluation, possible blood draws and empiric antibiotics they did not need, interrupted sleep, and anxiety — and their nurses absorb the workload.
5. **False negatives.** With sensitivity ≈ 65%, roughly one in three sepsis cases receives no alert. If clinicians come to rely on the system, a silent screen may read as reassurance — the automation-bias path to a missed case.
6. **System effects.** Dozens of alerts per unit per week. External validation of a widely deployed proprietary sepsis model found performance well below the vendor's published figures, with a large alert burden — the recipe for alert fatigue.

The dispute

Hospital leadership proposes activating SepsisGuard on every admitted patient, with no patient-level opt-out, disclosed only by one line in the admission packet. Their argument: the tool is population-level safety infrastructure — like a fire alarm — whose value depends on universal coverage; per-patient consent during acute admissions is infeasible and would produce inequitable monitoring.

A patient-advocacy coalition argues that a system which pages clinicians to the bedside, draws blood, and shapes treatment on the basis of an algorithm's probabilistic guess is not furniture — it is care. Patients are owed real disclosure and a meaningful way to say no.

Terms You'll Need

Read before Round 1. Each definition is written for readers with no AI coursework.

Probabilistic prediction

The model outputs a likelihood, not a diagnosis. A score of 85 means “elevated statistical risk based on patterns in past patients.” Some high-scoring patients are fine; some low-scoring patients are septic.

Sensitivity and positive predictive value (PPV), in plain terms

Sensitivity 65%: of every 100 patients who truly develop sepsis, the system flags 65 and misses 35. PPV 25%: of every 100 alerts fired, about 25 are real and 75 are false alarms. Both numbers matter, and they trade off as the alert threshold moves.

Alert fatigue

When most alarms are false, humans learn to tune them out — and eventually miss true ones. The classic example: EHR drug-interaction pop-ups, which clinicians override in the large majority of cases, sometimes reflexively. A low-PPV sepsis alert risks the same fate.

Automation bias

Over-trusting the machine: accepting its output in place of your own assessment, or feeling falsely reassured when it stays silent.

Model drift and external validation

A model that performed well where and when it was built may perform worse at your hospital, on your patients, years later. External validation tests it in new settings; drift is the slow decay as the world changes under it.

Opt-in vs. opt-out — and why defaults matter

Opt-in: patients must actively agree before inclusion. Opt-out: everyone is included unless they object. Decades of behavioral evidence show defaults powerfully determine participation — most people stay wherever the default puts them. Those who do opt out tend to differ systematically (health literacy, language, trust in institutions), so an opt-out design can quietly change who is protected.

Student Instructions

7. **Round 1 (6 minutes).** You are assigned FOR the hospital's policy (universal deployment, no opt-out) or AGAINST it (meaningful disclosure with an opt-out pathway). Argue your assigned position with the strongest case you can build. Do not equivocate.

Round 2 (4 minutes). Switch sides. Argue the position you just opposed, adding at least one argument your partner did not use. Do not equivocate.

8. **Round 3 (8 minutes).** Drop your assignments. Name the strongest surviving argument on each side, then design a third option that takes both seriously. Be specific: disclosure language, opt-out mechanics (if any), alert threshold, monitoring, re-validation schedule.

Round 4 (12 minutes). Class debrief anchored to the four principles, then the individual debrief grid below.

Round 4 — Four-Principle Debrief

The facilitator leads a short discussion anchored explicitly to the four principles, then each student completes the grid individually as an exit ticket. Grids are collected and scored — the session's structured evidence that students can define and apply all four principles to a live case.

Discussion anchors

9. **Beneficence.** What is the strongest evidence-based case that SepsisGuard helps patients — and what must be true of clinician response for the benefit to materialize?
10. **Non-maleficence.** Trace one concrete harm pathway for each error type: a false positive, a false negative, an ignored true alert. Which is the design's responsibility, and which the clinician's?
11. **Autonomy.** What exactly is the patient owed — information, agreement, an exit? Does one line in an admission packet meet any defensible disclosure standard?
12. **Justice.** Compare the equity effects of universal deployment versus opt-out. Who is protected, who is burdened, who ends up unmonitored under each design?
13. **Closing (tool, not team member).** Leadership called the system “safety infrastructure, like a fire alarm.” Where does the analogy hold, where does it break — and who answers for the missed case, whatever the model said?

Debrief grid (exit ticket)

For the third-way policy your pair designed in Round 3, write one to two sentences per principle: how your policy honors it — or knowingly trades it off, and why.

Principle	How our policy addresses it
Beneficence	
Non-maleficence	
Autonomy	
Justice	

Scoring rubric (0–2 per principle; 8 points)

14. **0 — Absent or misapplied.** Principle not addressed, or defined incorrectly (e.g., treats autonomy as the right to demand any treatment).
15. **1 — Named.** Correctly defined and invoked, but generically — the sentence could attach to any AI case.
16. **2 — Applied.** Correctly defined and connected to case-specific features — PPV and alert burden, opt-out equity effects, disclosure mechanics, error-type harms — with honest acknowledgment of trade-offs.

Suggested competence threshold: 6/8 with no zero. Below-threshold grids identify students for brief follow-up; item-level patterns identify weaknesses in the session itself.

Facilitator Guide

Anticipated arguments — FOR universal deployment

17. Population benefit depends on coverage; those most likely to opt out (or be steered out) may be at highest risk — opt-out can create the very inequity its advocates fear.
18. Per-patient consent during acute admission is unrealistic, and consent obtained that way would be low-quality anyway.
19. The EHR already runs invisible decision-support layers (interaction alerts, lab flags) that never required per-encounter consent.
20. The tool does not act on patients — clinicians do, and every downstream test and treatment still passes through ordinary consent.

Anticipated arguments — AGAINST

21. Autonomy is not contingent on operational convenience; “consent is hard here” has historically been the preamble to paternalism.
22. The fire-alarm analogy fails: a fire alarm doesn't draw your blood. At PPV ≈ 25%, the system routinely converts guesses into needles, antibiotics, and lost sleep for patients who were never asked.
23. A deterministic safety check differs in kind from a probabilistic prediction with documented bias, drift, and validation failures.
24. Trust is the long game: patients who discover undisclosed algorithmic monitoring after a public failure will trust the institution less.

Facilitation notes

25. Pair students; assign positions by counting off; enforce timing with a visible timer.
26. Interrupt pairs that converge early — the learning is in inhabiting the assigned side.
27. Watch for one student carrying both rounds; coach the quieter partner into Round 2.
28. Reject vacuous syntheses (“do both”): require specifics — disclosure wording, opt-out mechanics, thresholds, review schedules.
29. **Resist declaring a winner.** The assessment target is quality of principle-based reasoning, captured individually by the grid.