AI Limitations and Biases

Active Learning

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

Format rationale: the session has five distinct sub-domains, each requiring its own expert understanding. Five expert groups develop expertise in one sub-domain each, then re-mix into home groups of five, one expert per sub-domain, to teach each other. The act of teaching is the consolidation mechanism.

Timing at a glance (75 minutes)

Phase	Time	What happens
0 — Orientation	5 min	Facilitator frames the task and assigns groups.
1 — Expert groups	20 min	Read the primer; work the primer's mini-case; build the 2-minute teach-out.
2 — Home groups	30 min	Five teach-outs (about 4 minutes each: 2 to teach, 1 clarifying question, 1 buffer).
3 — Application round	12 min	Home groups work the shared vignette; each expert must contribute their lens.
4 — Synthesis & debrief	8 min	One-sentence synthesis per home group; facilitated debrief.

Two implementation tiers

- **Tier A (high-tech).** Expert groups may additionally pull up the primary source cited in their primer (all are indexed in PubMed or freely available) and read the abstract together.
- **Tier B (low-tech).** Printed primers only. The primers are written to be self-sufficient; the two tiers are functionally identical for running the activity.

Student facing instructions

You will be assigned to one of five expert groups. Your group becomes the class's expert on one failure mode of clinical artificial intelligence. You will then create a new group of five — one expert per failure mode — and teach the other four.

Phase 1 — Expert group (20 minutes)

- Read your assigned one-page primer individually (about 5 minutes).
- Work the mini case at the bottom of your primer as a group (about 7 minutes). It has no single right answer; it exists to make you discuss the material.
- **Build your 2-minute teach-out (about 8 minutes).** It must include: (a) a one sentence definition in your own words; (b) one concrete example, the primer's, or a better one; (c) one mitigation a future physician can actually do; (d) one question you would ask a vendor or your institution about this risk.

Phase 2 — New group (30 minutes)

One expert from each of the 5 previous groups joins a new group. Each expert teaches in turn — two minutes, then one clarifying question from the group. A timekeeper holds the schedule; every expert must finish. Do not read your primer aloud; teach from your own notes.

Phase 3 — Application round (12 minutes)

Your facilitator will hand out a short vignette that touches all five failure modes. As a home group, identify where each failure mode appears (or could appear) and name the single mitigation you would prioritize. Each expert is responsible for catching their own failure mode and for saying so out loud.

Phase 4 — Synthesis (8 minutes)

As a home group, write ONE sentence that captures the unifying lesson across all five failure modes. Specific beats general: "AI has problems" will be sent back for revision.

Facilitator guide

FACILITATOR PRE-TRAINING (30 minutes, before the session)

1. Read all five primers (15 min). You do not need depth beyond them; every claim a student should make is grounded in a primer. Note the one factual correction embedded in Primer A: it is not accurate that no current model can ground its claims — retrieval-augmented systems exist; the accurate claim is that self-confirmation is not verification.

2. Read the application vignette and its key (5 min). Both are printed below. The key lists where each failure mode hides; you arbitrate disputes but should let groups argue first.

3. Walk the logistics (5 min). Decide group assignments in advance (count the roster; see scaling below). Prepare five primer stacks and one vignette stack. Post the phase timings where everyone can see them.

4. Rehearse the two hard moments (5 min). Minute 35: if any home group is still on its second expert, call a hard reset to expert three. Phase 3: groups will want the facilitator to adjudicate immediately — hold back for at least four minutes.

RUNNING THE ACTIVITY

Class-size scaling. Ideal 25–30 (five expert groups of 5–6, then home groups of 5). Larger classes: run two parallel jigsaws with identical materials. Smaller classes (<15): reduce to three failure modes — combine Primers A+E (fabrication & accountability) and B+C (bias & drift), keep D standalone and shorten Phase 2 to 20 minutes.

Educational purpose. The Jigsaw activates the analyze/evaluate objectives for each sub-domain: students must understand their domain well enough to teach it, then apply all five lenses jointly in Phase 3.

Anticipated responses. Expert groups vary in confidence. Group E (liability) often feels the topic is "not for them yet"; re-anchor: this framework decides what they will be expected to verify in their first month of internship. Group C sometimes gets lost in statistics; the primer deliberately needs none beyond internal/external.

Watch for. (1) Experts finishing early and disengaging, have them draft one exam-style question on their topic. (2) One expert dominating a home group, enforce the timekeeper role. (3) Synthesis sentences that are too generic, require a mechanism ("every failure mode ends with a clinician trusting an output they did not verify") rather than a mood.

Common pitfalls. Faculty delivering mini-lectures during Phase 2. Resist: the pedagogical value is students teaching students. Intervene only for outright factual error, and even then by question ("does the primer agree with that?").

Closing debrief prompts (choose two). "Which failure mode surprised your home group most, and why?" • "Which failure mode are your future patients most at risk from in the next five years?" • "What is one verification or workflow habit each of you will commit to before clerkships?"

Application vignette (Phase 3 handout)

VIGNETTE — "A Tuesday on the wards" (student handout; print separately)

An intern admits Mr. Duran, 66, with pneumonia, overnight. To save time she asks a general purpose chatbot to summarize the evidence for adding a second antibiotic; it replies with a fluent paragraph citing two trials. The EHR shows a deterioration risk score of "low"; the score was validated at the health system's flagship hospital across town three years ago, before this community hospital switched laboratory vendors. At 03:00 the bedside monitor's early warning alert fires for the fourth time tonight on the same patient; the night nurse silences it within seconds, as she has the previous three. At 06:45 the intern,

reassured by the low risk score, presents Mr. Duran as "stable, low risk" — she has not re-examined him since midnight. On rounds, the attending asks her to explain how the risk score works. She cannot but says, "the model is FDA-cleared, so we're covered if anything happens."

FACILITATOR KEY — where each failure mode lives

A — Fabrication. The chatbot's two cited trials are unverified; nothing in the vignette says anyone checked them. Priority mitigation: verify both citations in PubMed before the evidence touches a treatment decision.

B — Bias. Subtlest thread — deliberately so. What population was the flagship-hospital model trained and validated on, and does this community hospital's population match? No subgroup performance has been reported. Strong groups also flag that "low risk" outputs get less scrutiny in exactly the patients whose data the model represents worst.

C — Drift / generalization. Validated at a different site three years ago; lab vendor since changed — two distinct problems (site transfer and dataset shift) students should name separately.

D — Automation bias & alert fatigue. The silenced alarm is fatigue-driven omission; the intern's "stable, low risk" without re-examination is commission — the score substituted for the bedside assessment. The Read-Record-Reconcile counterfactual: what would her presentation contain if she had committed her own impression first?

E — Liability & accountability. "FDA-cleared, so we're covered" is the target misconception: clearance does not transfer responsibility, and the clinician who acts on (or is reassured by) the output remains the accountable decision maker. Her inability to explain the score is the explainability thread.

Adjudication note. Award the point to any group that catches D twice (nurse and intern) — the double appearance is intentional.

The five expert primers (Appendices A–E)

Primer A — Hallucinations & persuasive fabrication

Definition. A hallucination is content a large language model (LLM) states fluently and confidently that is not grounded in any real source. Persuasive fabrication is hallucination dressed in authoritative form a complete citation, correct sounding mechanism, expert vocabulary, so the surface form recruits your trust.

Mechanism. An LLM generates tokens using learned conditional probabilities; decoding and system design affect the output. When the most plausible looking continuation is a citation that was never written, the model produces it with the same fluency as a real one; fluency alone does not reliably separate the two. Asking the model "are you sure?" re-runs generation, not verification: the plausible continuation of a challenge is confirmation plus new invented detail. Some systems now retrieve real documents before answering (retrieval-augmented generation, RAG). RAG can reduce fabrication and can show sources but it can retrieve the wrong or outdated source, and can misstate what a real source says. So verification shifts to content accuracy; it does not disappear.

Documented example, 2023. Asked 20 medical questions with supporting references, ChatGPT supplied 59 citations of which 41 (69%) were entirely fabricated with real-sounding authors, plausible titles, and credible journal formatting (Gravel J, et al. Mayo Clin Proc Digit Health 2023;1(3):226–234).

Mitigations. (1) Verify every AI supplied citation in PubMed or the journal archive before use, existence first, then whether it says what the model claims. (2) Cross-check clinical claims against a tertiary source or the issuing specialty society. (3) Prefer tools that cite retrievable sources and still open the sources. (4) Never accept the model's self-confirmation as a check.

One vendor question. "What is this tool's measured fabrication rate on clinical queries, how was it measured, and does the tool show its retrieved sources for inspection?"

MINI-CASE (work as an expert group)

Your classmate's study guide, built with an LLM, states a drug interaction you have never heard of, with a citation. She says: "I asked it twice and it gave the same answer both times, so it's solid." What exactly is wrong with her verification method and what would a real check look like, in under five minutes?

Primer B — Algorithmic bias & health equity

Definition. Algorithmic bias is systematic error that disadvantages identifiable groups. In clinical AI it most often enters through proxy bias: the model is trained to predict a measurable label that stands in for the true clinical goal, and the proxy carries historical inequity into the model even when race or ethnicity is never an input.

Mechanism. A model never predicts "health need" directly; it predicts the label you chose. In the canonical example, an algorithm used to allocate extra care predicted future healthcare costs as a proxy for future health needs. Because less money had historically been spent on Black patients at the same level of illness, equal predicted cost meant unequal actual sickness and the algorithm systematically under-identified Black patients for extra care (Obermeyer Z, et al. Science 2019;366:447–453). Underdiagnosis was also documented in chest-radiograph AI across underserved groups. The study does not establish historical under-documentation as the single cause of those disparities (Seyyed-Kalantari L, et al. Nat Med 2021;27:2176–2182).

It can be invisible. Deep-learning models can predict a patient's self-reported race from medical images with high accuracy, a task no human can perform so a model can quietly use race as a shortcut in ways clinicians cannot inspect (Gichoya JW, et al. Lancet Digit Health 2022;4:e406–e414). Language models add a cultural layer: GPT-4, tested across clinical tasks, produced differentials and assessments that tracked race and gender rather than the clinical facts (Zack T, et al. Lancet Digit Health 2024;6:e12–e22).

Mitigations. (1) Ask what the model actually predicts — is the target a proxy, and for whom does the proxy fail? (2) Demand performance reported by subgroup, not just overall. (3) Ask whether any race adjustment is doing clinical work without biological justification. (4) Support bias audits before and during deployment.

One vendor question. "Show me sensitivity and false-alarm rates broken out by race, ethnicity, sex, language, and insurance status, on data from a population like ours."

MINI-CASE (work as an expert group)

Your hospital is considering a model that predicts which discharged patients should get follow-up phone calls. It is trained to predict 30-day emergency-department return visits. A committee member says: "Return visits are objective, no bias possible." As the equity expert: what could the ED-return label be a proxy for, which patients might it under-serve, and what one report would you demand before deployment?

Primer C — Model drift & the generalization gap

Definition. The generalization gap is the difference between a model's performance on data like its training data (internal validation) and its performance in a genuinely different place, population, or time (external validation). Model drift (dataset shift) is the time version: accuracy decays as the world that produces the inputs changes, with no change to the model's code.

Mechanism. Models learn the statistical fingerprint of their development data — its patients, machines, assays, and documentation habits. Change any of these (new site, new scanner, new lab vendor, new

coding templates, shifting case mix) and the fingerprint no longer matches. Nothing crashes; the numbers just quietly stop meaning what they meant.

Documented examples. A widely deployed proprietary sepsis model with strong vendor-reported internal metrics showed markedly worse discrimination and missed most clinician-identified sepsis cases in external validation at a deploying hospital (Wong A, et al. JAMA Intern Med 2021;181:1065–1070). A diabetic-retinopathy screening system with specialist-level laboratory accuracy, deployed across 11 real clinics, rejected about a fifth of images as ungradable — clinic lighting and cameras differed from the pristine development images (Beede E, et al. Proc CHI 2020). And acute-kidney-injury models tracked across nine years drifted into steadily worse over-prediction as the underlying data shifted (Davis SE, et al. J Am Med Inform Assoc 2017;24:1052–1061).

Mitigations. (1) Ask whether a tool was externally validated — ideally on your patients, devices, and workflow — before trusting it. (2) Treat the training date like an expiration concern; ask when performance was last re-checked. (3) Expect institutional monitoring with recalibration triggers, and re-validation after any change in equipment, lab methods, or documentation practice.

One vendor question. "When and where was this model last externally validated, what happened to performance, and what monitoring runs after deployment?"

MINI-CASE (work as an expert group)

A vendor demo shows a deterioration-prediction model with an internal AUROC of 0.91. Your hospitalist asks: "Isn't 0.91 simply good, wherever it runs?" As the drift expert: explain in plain language why the number may not travel — name at least three specific things about your hospital that could break it — and say what evidence would change your mind.

Primer D — Automation bias, alert fatigue & the human-in-the-loop

Definition. Automation bias is the tendency to over-trust an automated output — deferring to it over your own judgment or the primary data. It has two textbook forms. Omission: the system stays quiet, so you stop looking, and miss what it missed. Commission: the system asserts something false, and you accept it over your own correct read. Alert fatigue is the workflow twin: drowned in low-value alerts, clinicians learn to dismiss everything — including the rare critical one.

Mechanism. The output arrives first and anchors you; scrutiny then feels redundant. The effect touches novices and experts alike — confidence in the source is not protection.

Documented examples. Residents reading electrocardiograms with the computer's interpretation attached did better when the computer was right — and worse than unassisted residents when it was wrong: incorrect advice actively pulled them off correct reads (Tsai TL, et al. J Am Med Inform Assoc 2003;10:478–483; automation bias reviewed in Goddard K, et al. JAMIA 2012;19:121–127). On the fatigue side, clinicians override 49–96% of drug-safety alerts (van der Sijs H, et al. JAMIA 2006;13:138–147).

The concrete defense — Read, Record, Reconcile. This session's named procedure, and your teach-out's centerpiece: (1) READ the primary data — image, labs, patient — before viewing any AI output. (2) RECORD your own impression in one written line; the written commitment is the mechanism, because an uncommitted impression silently bends toward whatever the model says. (3) RECONCILE: open the output, treat any discrepancy as a finding to investigate, and document your reasoning either way. A quiet model gets the same treatment as a loud one — silence is an output too.

Mitigations (system level). Tune alert specificity; tier by severity; suppress duplicates; track override rates as a patient-safety signal rather than a clinician failing.

One vendor question. "What is the alert's positive predictive value in deployed use, and what is the observed override rate at institutions like ours?"

MINI-CASE (work as an expert group)

A night-float resident tells you: "I always glance at the AI read first — it's faster, and I'd catch anything major." Using the ECG study, explain the specific risk in her workflow, then walk her through Read-Record-Reconcile on tonight's next film — including what she should write down, and when.

Primer E — Liability, explainability & the standard of care

Definition. Under one useful but context-dependent legal framing, the clinician functions as the learned intermediary between an AI tool and the patient: the person who interprets and acts on the output is the responsible decision-maker. Acting on AI advice or using a cleared tool does not automatically remove the clinician's duties. Institutions and vendors may also have responsibilities, depending on the facts and jurisdiction.

Where liability concentrates. Mello and Guha (N Engl J Med. 2024;390:271–278) analyze several possible liability scenarios: AI offers the most protection when it agrees with the standard of care and you follow it; exposure is greatest when you follow AI advice that deviates from the standard of care — or reject a correct, standard-concordant recommendation without documented reasoning. Doctrine is evolving, and none of this is legal advice; the stable core is that clinicians retain professional duties, alongside possible institutional and vendor responsibilities. (Concise earlier treatment: Price WN, et al. JAMA 2019;322:1765–1766.)

Explainability. Informed consent rests on telling a patient, in broad strokes, why a course of action is recommended. Opaque outputs complicate that — and current "explainable AI" methods can produce plausible explanations that do not faithfully reflect what the model did (Ghassemi M, et al. Lancet Digit Health 2021;3:e745–e750). Be cautious acting on a recommendation you cannot explain.

A tool, not a team member. AI is not a licensed professional and does not assume a clinician's professional duties. It can inform a decision; it cannot own one. Framing a tool as a "colleague" invites misplaced trust and blurs responsibility — document your reasoning, not just the model's output.

Mitigations. (1) Before acting on an output, ask: could I explain this recommendation to the patient, and would I make this decision without the tool? (2) Document your independent reasoning, especially when you follow or override AI against the grain. (3) Know your institution's policy on which tools are sanctioned for what.

One vendor question. "What does the vendor contractually represent about performance — and what does the license say about responsibility when the tool is wrong?"

MINI-CASE (work as an expert group)

An attending says on rounds: "If the sepsis model misses a case, that's on the hospital that bought it, not on me." As the accountability expert: sort out what is right and wrong in that sentence — separating vendor product liability, institutional responsibility, and the clinician's own duty — and state what you would want documented in the chart when you override a model.

Sources cited in the primers

- **Gravel J, D'Amours-Gravel M, Osmanlliu E.** Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health. 2023;1(3):226–234.
- **Obermeyer Z, Powers B, Vogeli C, Mullainathan S.** Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453.
- Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of AI algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27:2176–2182.

- **Gichoya JW, Banerjee I, Bhimireddy AR, et al.** AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health. 2022;4(6):e406–e414.
- **Zack T, Lehman E, Suzgun M, et al.** Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12–e22.
- **Wong A, Otles E, Donnelly JP, et al.** External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065–1070.
- **Beede E, Baylor E, Hersch F, et al.** A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proc 2020 CHI Conf Hum Factors Comput Syst. 2020:1–12.
- **Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME.** Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052–1061.
- **Tsai TL, Fridsma DB, Gatti G.** Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc. 2003;10(5):478–483.
- **Goddard K, Roudsari A, Wyatt JC.** Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121–127.
- **van der Sijs H, Aarts J, Vulto A, Berg M.** Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138–147.
- **Mello MM, Guha N.** Understanding liability risk from using health care artificial intelligence tools. N Engl J Med. 2024;390(3):271–278.
- **Price WN 2nd, Gerke S, Cohen IG.** Potential liability for physicians using artificial intelligence. JAMA. 2019;322(18):1765–1766.
- **Ghassemi M, Oakden-Rayner L, Beam AL.** The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–e750.