MONASH University

WEEK 3 CLASS ACTIVITY — AI ETHICS REVIEW BOARD

Trust, Ethics & Governance · LO5, LO3 · ~60 minutes

> **Goal.** Run the lecture's five-question **trust scorecard** on a real AI system, then on your own idea — and propose fixes. Your team is an ethics review board: your job is to find the problems *before* they reach real people.

You work in teams of 3–4 (cross-campus teams via the Google Meet breakout). Project deck: `activity_slides.pptx`.

THE TRUST SCORECARD (YOUR REFERENCE)

#	Question	Score it on…
1	**Bias & fairness**	Does it treat people equally across groups (gender, race, age, region)?
2	**Transparency**	Do people know they're dealing with AI / that AI shaped a decision?
3	**Traceability**	Can you explain *why* a decision was made — and reconstruct it later?
4	**Privacy & data**	Is data collection minimal, consented, protected, and not over-retained?
5	**Security**	Safe from others (breaches) — *and* not predatory (surveillance creep)?

Score each 🟢 **green** (ready), 🟡 **amber** (fix it), or 🔴 **red** (not ready).

PART A — AUDIT THE CASE: "FACEENTRY" (≈25 MIN)

> **The case.** A university's leadership wants facial recognition to take class attendance automatically. Cameras at the lecture-hall door scan each student's face, match it against a photo database of all enrolled students, log attendance, and flag anyone "absent" to their tutor.

As a board, score FaceEntry on all five questions. Write **one line of reasoning** per row.

MONASH University

#	Question	🟢 / 🟡 / 🔴	One-line reason
1	Bias & fairness		does it recognise every student equally well?
2	Transparency		do students know they're scanned? did they consent?
3	Traceability		if it marks you absent wrongly, can you contest it?
4	Privacy & data		biometric data of every student — where is it stored?
5	Security		a face database is a breach target — and is this surveillance creep?

There's no single "right" score — defend your reasoning. (Hint: at least one row is a clear red.)

PART B — AUDIT YOUR IDEA (≈20 MIN)

Run the **same five questions** on your own Assessment 1 idea. Three rules:

- **Be honest** — the point is to *find* your reds, not look good. A red you've spotted is a red you can fix.
- **Be specific** — "AI can be biased" scores nothing. "Our model has seen fewer informal businesses, so it may rate them worse" is a finding.
- **Most ideas have ≥1 red** — if you scored all green, you're not looking hard enough. Push each other.

#	Question	🟢 / 🟡 / 🔴	Why — specific to *your* idea
1	Bias & fairness		
2	Transparency		
3	Traceability		
4	Privacy & data		
5	Security		

PART C — MITIGATE (≈15 MIN)

For **every red or amber**, propose a concrete fix using the **three moves**:

1. **Bound the blast radius** — narrow the AI's job so a mistake does small, recoverable harm (assist, don't decide).
2. **Human in the loop** — a qualified human reviews / signs off / can override the high-stakes calls.
3. **Document & audit** — log decisions, data, model version, and the choices you made, so you can explain it later.

> Write: *"Risk: [the red] → Move: [1/2/3] → Mitigation: [what you'll actually do]."*

MONASH University

DEBRIEF (≈10 MIN, WHOLE CLASS)

- Each board shares the **one red** that worried them most, and the move they'd use to fix it.
- **The lesson:** a red you *name and mitigate* scores **higher** than a product that pretends to be all green. Honest risk-thinking is exactly what Assessments 1 and 3 reward.
- **Keep your filled scorecard** — it's the risks section of your Assessment 1.

FACILITATOR NOTES (DO NOT DISTRIBUTE)

- **FaceEntry answer sketch:** Bias 🔴 (face recognition is less accurate for some groups → wrongful "absent"); Transparency 🟡/🔴 (constant scanning without genuine consent); Traceability 🟡 (need a contest/appeal path + logs); Privacy 🔴 (biometric data of all students is highly sensitive under PDPA; storage/retention unjustified); Security 🔴 (a biometric database is a prime breach target *and* normalises surveillance). The teaching point: a "convenient" feature can be a governance minefield — and the fix is often to **not** collect biometrics at all (bound the blast radius: a QR check-in does the same job with none of the risk).
- **Facilitate, don't lecture.** Push with "which group does it fail, exactly?" and "where does that data live?" Let teams discover the reds.
- **Cross-campus:** one shared scorecard per mixed team (Jamboard / shared doc); Lead B keeps the remote campus scoring, not just watching.
- **Bridge:** end by pointing every student at the Week 3 post-class task (write the risks & mitigations section) and the Assessment 1 brief (due end of Week 4).