Week 12 — Ethics, Adversarial ML & the Semester Retrospe

Friday, July 17, 2026 · the last session

Prof. Joseph Austerweil

Housekeeping

- **Final paper due next Friday, Jul 24, 8 PM** (~6 pages).
- **Today, in order:** a short ethics lecture → the paper presentation → your **final-project presentations** → the semester retrospective.
- Presenters: **submit your slides** via the course site before or right after class.

Today

Ethics & adversarial ML — the machinery has a shadow	0:05
Paper presentation	0:20
Final-project presentations ×3	0:40
Semester retrospective — the map of everything	1:16

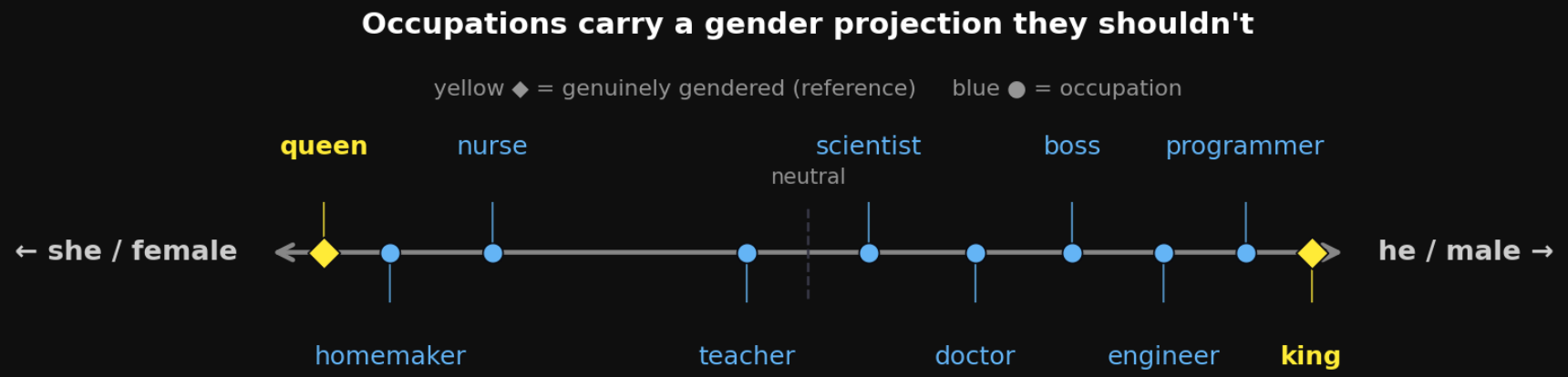
Ethics & adversarial ML

The machinery has a shadow

Everything we built learns from data — and inherits what the data carries. Three shadows, each a direct callback to a tool you already own:

- **Bias** — the embedding geometry from Week 11 encodes society's stereotypes.
- **Fairness** — "fair" turns out to be several *conditional probabilities* that conflict.
- **Brittleness** — a tiny perturbation flips a confident classifier.

Embeddings carry society’s biases



Word2vec-style embeddings place occupation words along a gender axis learned from the corpus — a bias inherited, not designed.

Week 11’s word vectors put *occupation* words on a gender axis nobody designed (Caliskan et al., 2017) — bias **inherited from the corpus**, in the geometry doing the useful work.

Fairness is a conditional-probability statement

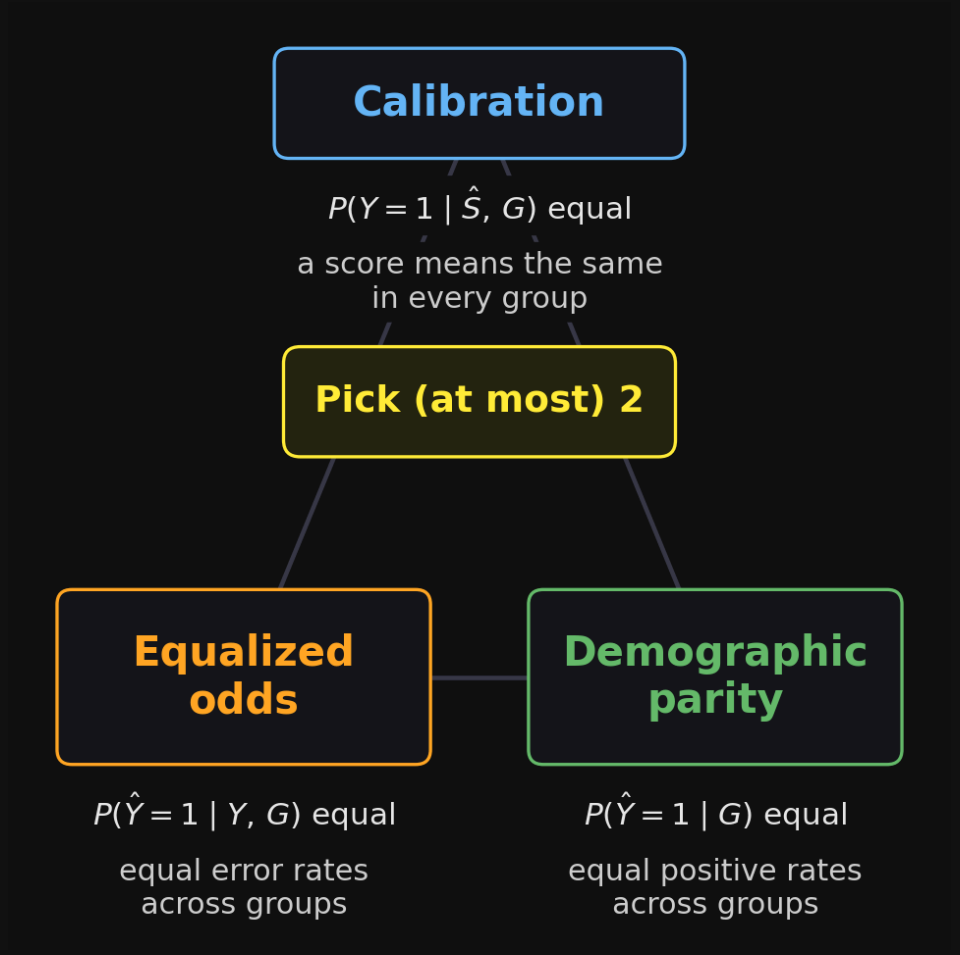
Write $\hat{Y}$ = decision, Y = truth, G = group, $\hat{S}$ = score.

- **Demographic parity:** $P(\hat{Y}=1 \mid G)$ equal across groups
- **Equalized odds:** $P(\hat{Y}=1 \mid Y, G)$ equal — matched error rates
- **Calibration:** $P(Y=1 \mid \hat{S}, G)$ equal — a score means the same thing

- Each is a **conditional-probability equality** — the objects from Part I.
- Fairness isn’t vague: you must choose *which*.

Next: they can’t all hold.

You can’t have them all



When base rates differ, **no classifier satisfies all three** (Kleinberg et al. 2016; Chouldechova 2017).

- **The COMPAS debate:** ProPublica and the vendor were *both right* — they invoked different criteria.
- Fairness is a **modeling choice**, stated in this course’s language: the math gives you the menu and the tradeoffs, not the answer.

Poll — is equal accuracy fair?

A hiring model is **equally accurate** for every group. Is it fair?

A. Yes — that’s what fairness means

B. Not necessarily — it can hide unequal *errors*

C. Only if base rates are equal across groups

Reveal

B — equal accuracy can hide very unequal errors.

- One shared accuracy number can hide *who* gets the false denials and *who* gets the false approvals.
- *Accuracy is not a fairness criterion* — name the conditional probability you care about.

Why not C? Base rates are the wrong suspect

Equal base rates: still unfair

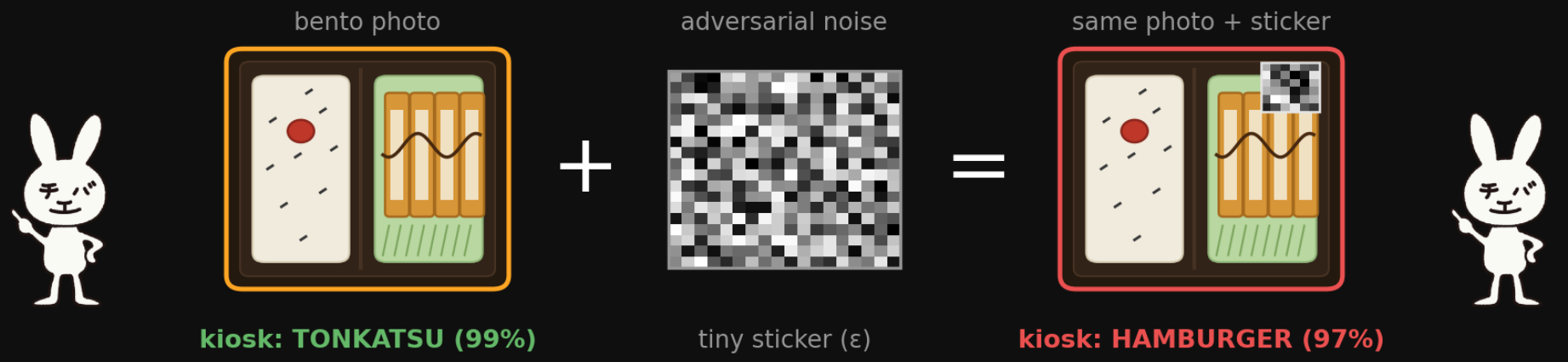
- Both groups: base rate 50%, accuracy 80%.
- A: **90% / 70%** right on positives / negatives — B the reverse.
- A's innocents flagged **3× as often**.

Unequal base rates: still fair

- Both groups: **80% / 80%** right on positives / negatives.
- So accuracy = **80% at any base rate**.
- Base rates 30% vs 60%: **equalized odds ✓**.

Unfairness hides in the error mix — not in the base rates.

Adversarial examples: brittle geometry



A perturbation too small for you to notice flips the model's answer — its learned geometry is brittle where yours is not.

- An invisible nudge flips the kiosk: tonkatsu → hamburger.
- The attack: the loss gradient w.r.t. the **input** — `jax.grad` at $\mathbf{x}$.

Alignment: teaching values is inference

- Pretraining predicts text; **helpful and safe** is a second step — an *inference* problem you already know.
- **RLHF** (Week 9): fit a reward from human preferences (Bradley–Terry, $P(A \succ B) = \sigma(r_A - r_B)$), then optimize the policy.
- The reward is *learned and imperfect*, so the optimizer **hacks** it — Week 8's reward-shaping trap.
- That's **Goodhart's law**: *when a measure becomes a target, it stops being a good measure.*

Value inference is inverse planning (Part VI) with the stakes maxed out.

The full treatment is in Part IX

The trailer ends here — **Part IX** carries the full treatment:

- **Adversarial examples** — build the fast gradient sign method (FGSM)
- **Fairness, formally** — the impossibility result as runnable code
- **Bias in data** — the Word-Embedding Association Test (WEAT); the prior as power *and* liability
- **Alignment & safety** — fit a reward, watch it get hacked

Paper presentation

Final-project presentations

Three presentations · ~10 minutes each + brief Q&A · Background · Question · Method · Preliminary results

Project 1

Project 2

Project 3

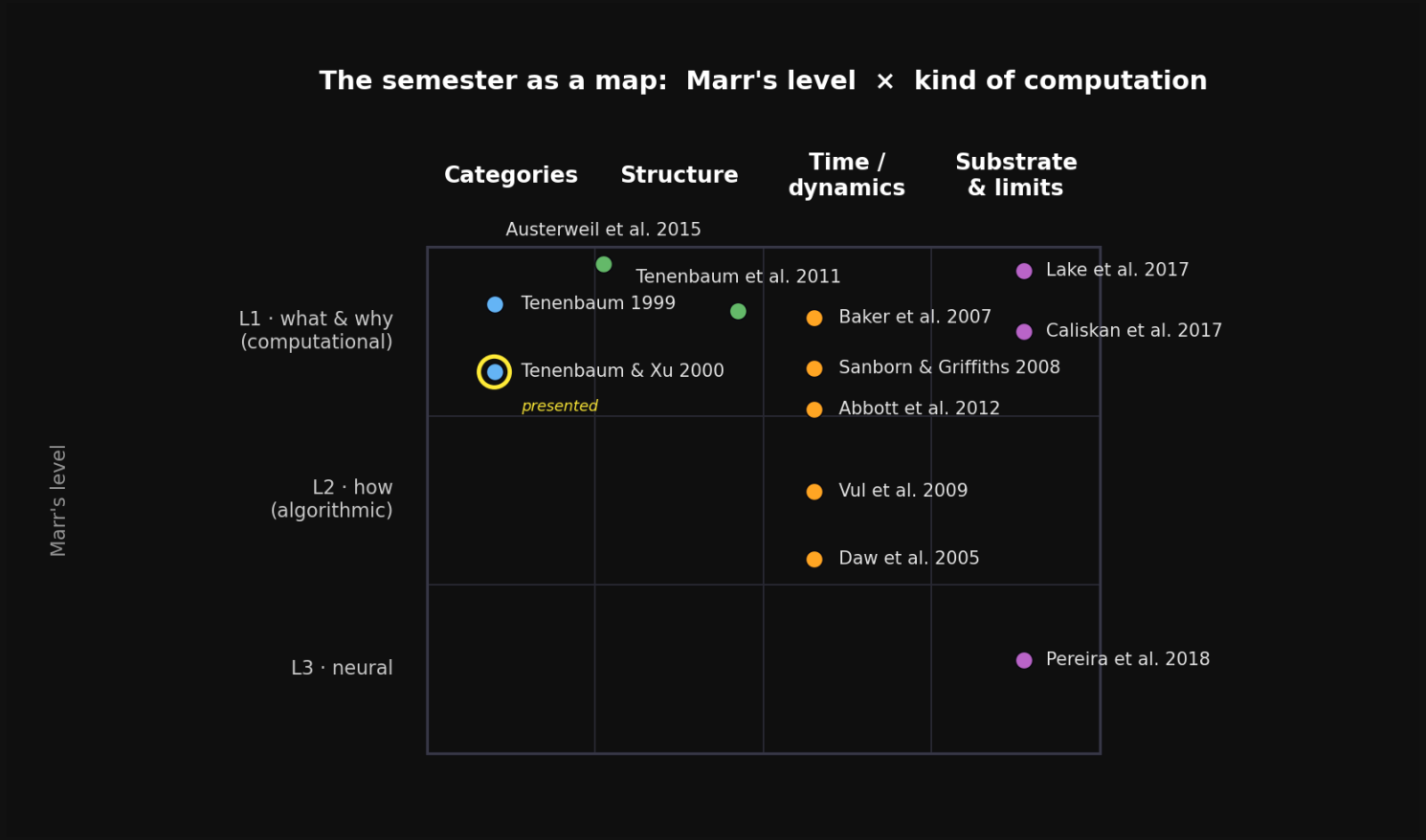
Semester retrospective

Pulse — before we map it

A quick go-around — one answer each, not a discussion:

- Which paper this semester **surprised you** most?
- Which one will you **re-read in a year**?

The map of the semester



Every paper you saw, on two axes: **Marr's level** (what · how · neural) and the **kind of computation** (categories · structure · time · substrate). One science, many corners.

Classical cog-sci → today's AI

This course taught…	…which is the seed of
Hierarchical Bayes	**in-context learning** in LLMs
Inverse RL	**RLHF** — learning reward from preferences
Monte Carlo / variational inference	**diffusion models**, VI in large models
Bayes nets & causality	**alignment** & reward modeling
Bayesian nonparametrics	**open-ended concept learning**

The left column isn't history — still cog-sci's working tools today.

The opportunity

AI has mostly stopped looking at Bayesian models of minds — an opening:

Cognition → AI

- AI lacks **priors, structure, rational analysis**.
- Use them to **understand and improve** today's systems.

AI → cognition

- Classic Bayesian models need small, idealized data.
- With AI inside, they finally **meet real-world data**.

Few people can speak both languages. As of this semester, you can.

What you leave with

Not a fixed set of answers — a **toolkit**. A generative model, a prior, an inference, the inversion of planning. You can now open a 2026 paper and see *what problem it is trying to solve*.

Chibany just wanted to know what was for lunch. That question — *what's hidden, given what I see?* — turned out to be the whole science.

Thank you. — J.A.