Human-Computer Interaction

Critical, Responsible & Sustainable AI

Professor Bilge Mutlu

Today's Agenda

- Topic overview: *Critical, Responsible & Sustainable AI*
- The frame: *enduring fundamental → how AI changes it*
- Provocation: *Design practice, governance, or refusal?*
- Two questions you asked: **Nov 30 crit** and **Dec 7 talks**
- Discussion

This term's frame

Enduring fundamental → How AI is changing it → Depth (a claim to argue with)

- **Fundamental:** accountability: what values a system serves, and who can check
- **Fault line: principles** anyone can sign vs. **mechanisms** someone can verify
- **AI:** scale, opacity, and hidden labor widen the gap; law has started to close it

This week's readings

- **Required**
  - Falco, Shneiderman et al., *Governing AI Safety through Independent Audits* (2021)
  - Birhane et al., *The Values Encoded in Machine Learning Research* (FAccT 2022)
- **Recommended:** Gray & Suri, *Ghost Work* (2019)
- **Optional:** Suchman (2007); Strubell, Ganesh & McCallum (ACL 2019)

The thread so far

Seminar	The question it left open
Sep 21 Human-AI	A guardrail that depends on willpower isn't one
Oct 5 Generative AI	Whose labor did the corpus launder?
Oct 26 Agents	"Autonomy" hides the human work that sustains it
Nov 9 Affective	Whose ground truth? Whose consent?
Nov 16 Privacy	The norm has to be applied *inside* the system

Today: who can hold the system, and the field, to account?

Topic overview: *Critical, Responsible & Sustainable AI*

Principles are not enough [1]

A survey of **84** sets of AI ethics principles found:

- **Transparency** in 73 · **justice & fairness** in 68 · **non-maleficence** in 60

"Whilst these principles are very worthy, they do not help practically with the design, development and use of automated systems."

Their proposal: independent audit of AI systems (**IAAIS**, pronounced "eyes")

[1] Falco, G., Shneiderman, B., Badger, J., et al. (2021). Governing AI Safety through Independent Audits. *Nature Machine Intelligence*, 3, 566–571.

The AAA principles [1]

Assessments

Name risks **before** deployment, for everyone the system touches

Audit trail

A **"black box"** like a flight data recorder: what the system did, in context

Adherence

Meet the rules of each **jurisdiction** and sector

[1] Falco, G., Shneiderman, B., Badger, J., et al. (2021). Governing AI Safety through Independent Audits. *Nature Machine Intelligence*, 3, 566–571.

Who makes anyone do it? [1]

- **Insurers:** require assessments and audit logs before they insure
- **Courts:** set the **standard of care**; who is liable when the system fails?
- **Government agencies:** incentives (tax breaks), or fines and shutdowns

A principle becomes accountability only when someone outside the company can check it.

[1] Falco, G., Shneiderman, B., Badger, J., et al. (2021). Governing AI Safety through Independent Audits. *Nature Machine Intelligence*, 3, 566–571.

Values are choices, not defaults [2]

100 most-cited **NeurIPS / ICML** papers, read line by line

- **68%** never mention a societal need; only **4%** rigorously justify one
- **98%** never mention a potential negative impact
- Top values: **performance** 96% · **generalization** 89% · **efficiency** 84% · **novelty** 77%
- **None** mention autonomy, justice, or respect for persons

---

[2] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *FAccT 2022*. doi:10.1145/3531146.3533083

"Technical" values are social [2]

- **Performance:** accuracy averaged with equal weight per example; groups rare in the data count less
- **Efficiency:** 84% value it, but only **15%** of those value using fewer resources; it means *scale up*
- **Novelty:** *technical* novelty; critiquing a harmful dataset doesn't count

These values, as applied, favor the already powerful.

[2] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *FAccT 2022*. doi:10.1145/3531146.3533083

Who writes the most-cited papers [2]

	2008/09	2018/19
Corporate-affiliated authors	24%	**55%**
Big Tech authors	13%	**47%**
Any corporate tie (authors or funding)	45%	**79%**

Of university-affiliated papers, **80%** came from top-50 universities.

[2] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *FAccT 2022*. doi:10.1145/3531146.3533083

The central tension: three poles

Where does responsible AI *live*? **Three poles, not a spectrum.**

Design practice

Change what the field values and builds *(Birhane)*

Governance

Audits, liability, binding law *(Falco)*

Refusal

Some systems shouldn't be built

But what about *AI*?

The labor behind "autonomy" [3]

The paradox of automation's last mile: automating a task creates new tasks for people.

- Labeling, moderating, checking: on-demand work, paid per task
- A five-year study of these workers in the **U.S. and India**
- The Oct 26 point again: "autonomy" hides the people who sustain it

[3] Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

Sustainability: the compute bill [4]

	CO_2e (lbs)
One passenger, NY ↔ SF flight	1,984
Average car, lifetime, incl. fuel	126,000
Transformer (big) + neural architecture search	**626,155**

Their asks: report training time and compute; fair access to compute; prefer efficient methods

[4] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *ACL 2019*.

Governance is now law [5]

EU AI Act (2024): obligations scale with **risk tier**

- **Unacceptable** (banned) · **high** · **limited** (transparency) · **minimal**
- High-risk: risk management, logging, human oversight, conformity assessment
- **Timing:** bans since Feb 2025; most high-risk duties now **Dec 2027**

Falco et al. cite the 2021 proposal; their AAA maps onto the law.

[5] Regulation (EU) 2024/1689 (the EU AI Act), amended in 2026 by the Digital Omnibus on AI. https://artificialintelligenceact.eu

What changes

- **Principles → audit rules** someone outside can check
- **"Neutral" metrics → stated values** you can be held to
- **"Autonomous" → name the labor** that keeps it running
- **"Efficient" → report the compute**
- **Guardrail:** *ethics-washing*: voluntary principles standing in for accountability

👉 The three poles are the instrument for reading any "responsible AI" claim.

Read your own paper like Birhane

Three checks for your draft, before Nov 30:

- **Justification:** does your intro name *whose* need this serves, and show the link?
- **Negative potential:** does your discussion name one harm or limit of your system?
- **Values:** which values does your abstract uplift? Only performance?

Your provocation (*post one slide*)

Pick one:

1. **Provocation:** where does responsible AI *live*: **design practice**, **governance**, or **refusal**? Defend one pole against the other two.
2. **Critical artifact:** take a real system or "responsible AI" pledge; show where it is accountability and where it is **ethics-washing**.
3. **Design / policy:** one build-time requirement (fairness, auditability, governance, sustainability) you would make *binding*, and how you would verify it.

Two questions you asked

Mon Nov 30 Studio Crit: what to bring

A three-slide deck, cast from a team laptop to your pod monitor:

1. **Question:** your research questions and hypotheses
2. **Analysis:** what you collected, from whom (*n*), and how you analyzed it
3. **Results:** your main result, reported the Nov 20 way, and the comparison your claim rests on

Team name and title go on slide 1. **Three slides, no more.**

Nov 30: how it runs

Same tables and cycle as the Nov 13 crit:

Time	What happens
0:00 · 8 min	Frame + set up at your table
0:08 · 42 min	**Crit:** per team, present 6 · silent writing 3 · round-robin 9 · commit 3
0:50 · 10 min	**Dry run:** each team gives its 2-minute talk; the table says what to cut
1:00 · 15 min	Debrief + Dec 4 / Dec 7 logistics

Mon Dec 7: Final Presentations

- **Lightning talks:** every team, **2 minutes**, back to back
- **Your deck:** one per team, on Canvas **before class**; all decks run from one laptop, so no setup
- **What's in it:** one question, one method, one finding, one implication: your Nov 30 deck plus an implication
- **No Q&A per team:** questions come at the end of the session
- **Then:** use the feedback in the final paper, due **Mon Dec 14, 11:59 pm**

Discussion

Today's roles

Reporter	Skeptic	Connector
Seat 5	**Seat 6**	**Seat 1**

Same in every pod. **Seats 2, 3 and 4** have no role today; you are still in the argument.

Discussion Format

In your pod, on your pod monitor: ~30 minutes.

- **Everyone shows their slide**, about 2 minutes each, cast from your own laptop
- **Skeptic** puts a counterargument on the table · **Connector** ties it to an earlier week
- **Reporter** writes the pod's three lines (agreed, split, unresolved) and delivers them to the room

What's Next?

- **Fri Nov 27:** no class (Thanksgiving)
- **Mon Nov 30 · Studio Crit:** your **three-slide deck** (question, analysis, results)
- **Fri Dec 4 · Writing & Presentation Clinic:** Reporting & Writing Deliverable due 9:30
- **Mon Dec 7 · Final Presentations:** 2-minute lightning talks