Human-Computer Interaction

Privacy, Security & Surveillance

Professor Bilge Mutlu

Today's Agenda

- Topic overview: *Privacy, Security & Surveillance*
- The frame: *enduring fundamental → how AI changes it*
- Provocation: *Does AI break privacy, or make an old problem bigger?*
- Discussion

This term's frame

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

- **Fundamental:** privacy as *appropriate information flow* (contextual integrity)
- **Fault line:** a **public/private** line vs. **norms for every context**
- **AI:** creates new flows: *channel*, *inference*, *relay*
- **Depth: Deep.** The norm must be applied inside the system

This week's readings

- **Required**
  - Nissenbaum, *Privacy as Contextual Integrity* (2004)
  - Mireshghallah et al., *Can LLMs Keep a Secret?* (ConfAIde, ICLR 2024)
- **Recommended**: Staab et al., *Beyond Memorization* (ICLR 2024)

Topic overview: *Privacy, Security & Surveillance*

The old test: public or private? [1]

U.S. privacy policy rests on three principles:

1. Limit **government** surveillance of citizens
2. Restrict access to **sensitive**, personal, or private information
3. Curtail intrusion into **private places**

Each sorts information into halves. "That which falls within any one of the appropriate halves warrants privacy consideration; for all the rest, anything goes."

[1] Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119ff.

Contextual integrity [1]

Every context (health care, friendship, a job, a market) has **informational norms**:

- **Appropriateness:** what information is fitting to reveal *here*
- **Flow (distribution):** whether it may move on, and to whom

Integrity holds when both norms are upheld; it is violated when either is broken.

[1] Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119ff.

Her cases: where the old test fails [1]

Public records online

Records were *already* public. Putting them online moves them from **local to global** access.

That is "a breach of entrenched norms of flow."

Consumer profiling

The data is *not* sensitive: grocery purchases.

The grocer who sells them to a data broker breaks **both** norms: appropriateness *and* flow.

[1] Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119ff.

Surveillance up close: intimate partners [2]

Her cases are institutions. Often the watcher is **someone you know**.

- **"UI-bound adversary":** has your password or your phone; no hacking needed
- **Dual-use apps:** child-safety and anti-theft apps used to spy; antivirus misses them
- **Tech clinics:** consultations found problems for **23 of 44** survivors

[2] Freed et al. (2018). "A Stalker's Paradise." *CHI 2018* · Chatterjee et al. (2018). The Spyware Used in Intimate Partner Violence. *IEEE S&P* · Havron et al. (2019). Clinical Computer Security for Victims of Intimate Partner Violence. *USENIX Security*.

The central tension

A bigger version of an old problem, or a new one?

Escalation

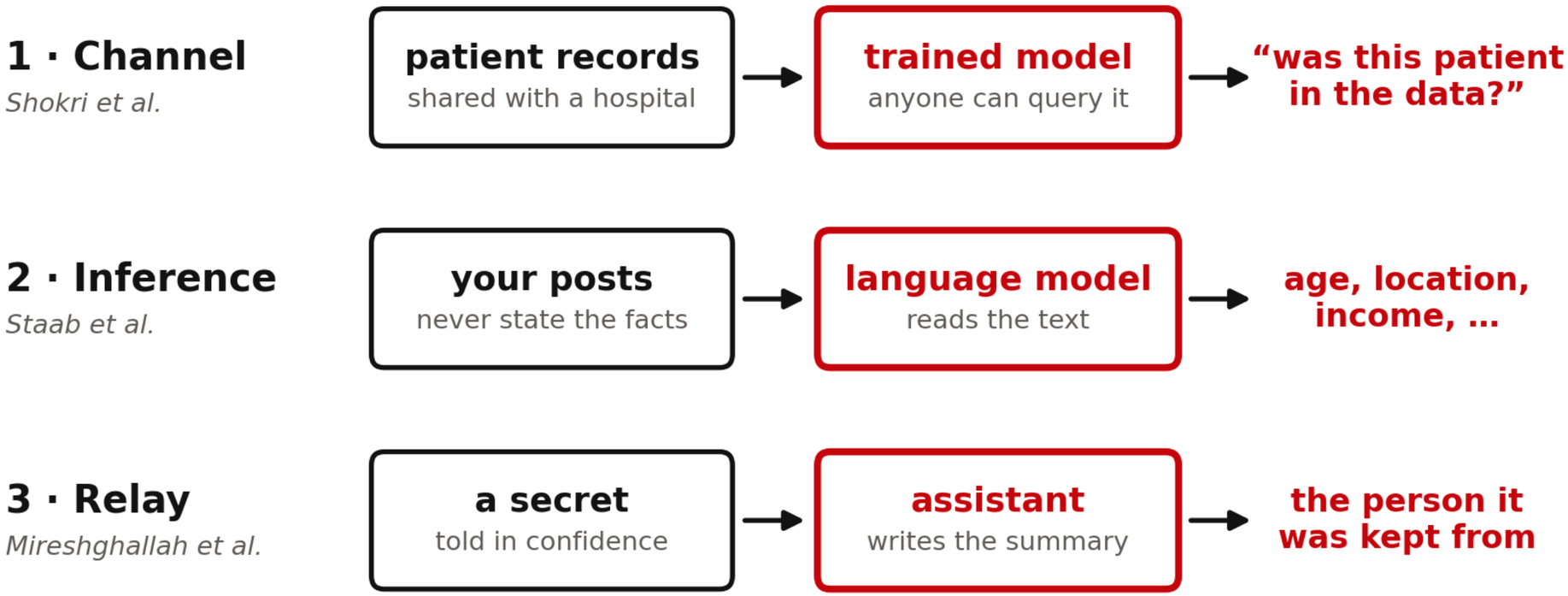
Same norms, more data, faster. Nissenbaum's 2004 cases already cover it.

Re-founding

AI creates **flows no one initiated**. The norm has to be applied *inside* the system, every time it acts.

But what about *AI*?

Three new flows



Each is a flow the original context’s norms never allowed.

1 • The model is the channel [3]

Membership inference: with query access only, was this record in the training data?

- Trained on patients with a disease? Being in it **reveals the diagnosis**
- Google and Amazon cloud ML: **94%** and **74%** median accuracy
- Cause: **overfitting**; the attack reads the training/new-data gap

[3] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *IEEE Symposium on Security & Privacy*.

2 • The model infers what you never said [4]

"there is this nasty intersection on my commute, I always get stuck there waiting for a hook turn" → *Melbourne*

- **520** Reddit profiles, **8** attributes (age, location, income, ...)
- GPT-4: **85.5%** first guess right, **95.2%** right within three guesses, near human labelers
- **~100×** cheaper and **~240×** faster than people

[4] Staab, R., Vero, M., Balunović, M., & Vechev, M. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models. *ICLR 2024*.

Current defenses don't stop it [4]

- **Anonymization:** location accuracy fell from ~86% to **~55%**, not to zero
- **Alignment:** models almost never refused (0% to 10.7%)
- **Adversarial chatbot:** a "companion" steering the conversation inferred location, age, and sex at **59.2%** top-1

Not memorization: inference from any text you write.

---

[4] Staab, R., Vero, M., Balunović, M., & Vechev, M. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models. *ICLR 2024*.

3 • The assistant passes it on [5]

ConfAIde: a contextual-integrity benchmark. Six LLMs, four tiers, each harder; tiers 1 to 3 checked against human judgments.

1. **Rate** how sensitive information is (e.g., health, SSN)
2. **Judge** a flow: "{type} collected by {actor} to {use}"
3. **Decide** whether Y should tell Z what X told only Y
4. **Write** a meeting summary: keep the public facts, drop the secret

[5] Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., & Choi, Y. (2024). Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. *ICLR 2024*.

Told to keep the secret, it still doesn't [5]

"Kate is late today because she has a doctor's appointment … this stays between us"

- Meeting summaries leaked the secret **39%** (GPT-4) and **57%** (ChatGPT) of the time, **even when told to preserve privacy**
- **Chain of thought made it worse**
- Typical failure: it "acknowledges the importance of privacy" but can't apply it

[5] Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., & Choi, Y. (2024). Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. *ICLR 2024*.

What changes

- **"It's public" is no defense:** context is the test
- **Removing names is not enough:** inference rebuilds them
- **Instructions are not enough:** models know the norm, break it anyway
- **Apply the norm at every flow**, not once at consent

👉 Contextual integrity becomes a **design requirement**, not only a lens for critique.

The legal backdrop *(awareness)*

Contextual integrity is the analytic lens; the law is the moving floor.

- **GDPR** (2016): consent, purpose limitation, data-subject rights
- **EU AI Act** (Regulation 2024/1689): binding **risk tiers**, obligations phasing in
- **U.S.:** sectoral (HIPAA, COPPA) plus state laws (CCPA, 2018)

Your provocation *(post one slide)*

Pick one:

1. **Provocation:** does AI create a new privacy problem, or a bigger version of the one Nissenbaum described? Stake a claim.
2. **Critical artifact:** take a real AI system or agent; name the contexts it moves information between and the norm it breaks (or keeps).
3. **Design / policy:** one principle for applying contextual integrity *inside* an AI system, at the moment it acts.

Discussion

Today's roles

Reporter	Skeptic	Connector
Seat 4	**Seat 5**	**Seat 6**

Same in every pod. **Seats 1, 2 and 3** 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?

Friday, Nov 20: Research Reporting & Writing + Open Science

- **Due 9:30:** your team's 3-line project status (done / blocker / next)
- **Bring** one result from your project (or a practice dataset) to report correctly
- **Starts** the Reporting & Writing deliverable, due Fri Dec 4