Human-Computer Interaction

Responsible & Ethical Design; Trust & Transparency

Professor Bilge Mutlu

Today's Agenda

- Last assignment
- Project final steps
- Course evaluation
- Topic overview: *Responsible & Ethical Design, Trust, & Transparency*
- Discussion

Last Assignment

Skills practiced:

» Data analysis (Nov 12 & 19 lectures)

» Reporting (Dec 3 lecture)

Due Friday Dec 5

Project Final Steps

Due dates:

» Data analysis — Dec 5

» Presentations — Dec 8 & 10 (more on this on Wednesday)

» Final paper — Dec 12

Course Evaluation

» Please complete student evaluations

» Search for "heliocampus" in your email

» Helps us know how the course supported your learning and how it can be refined, and helps us make the case to the department for similar courses

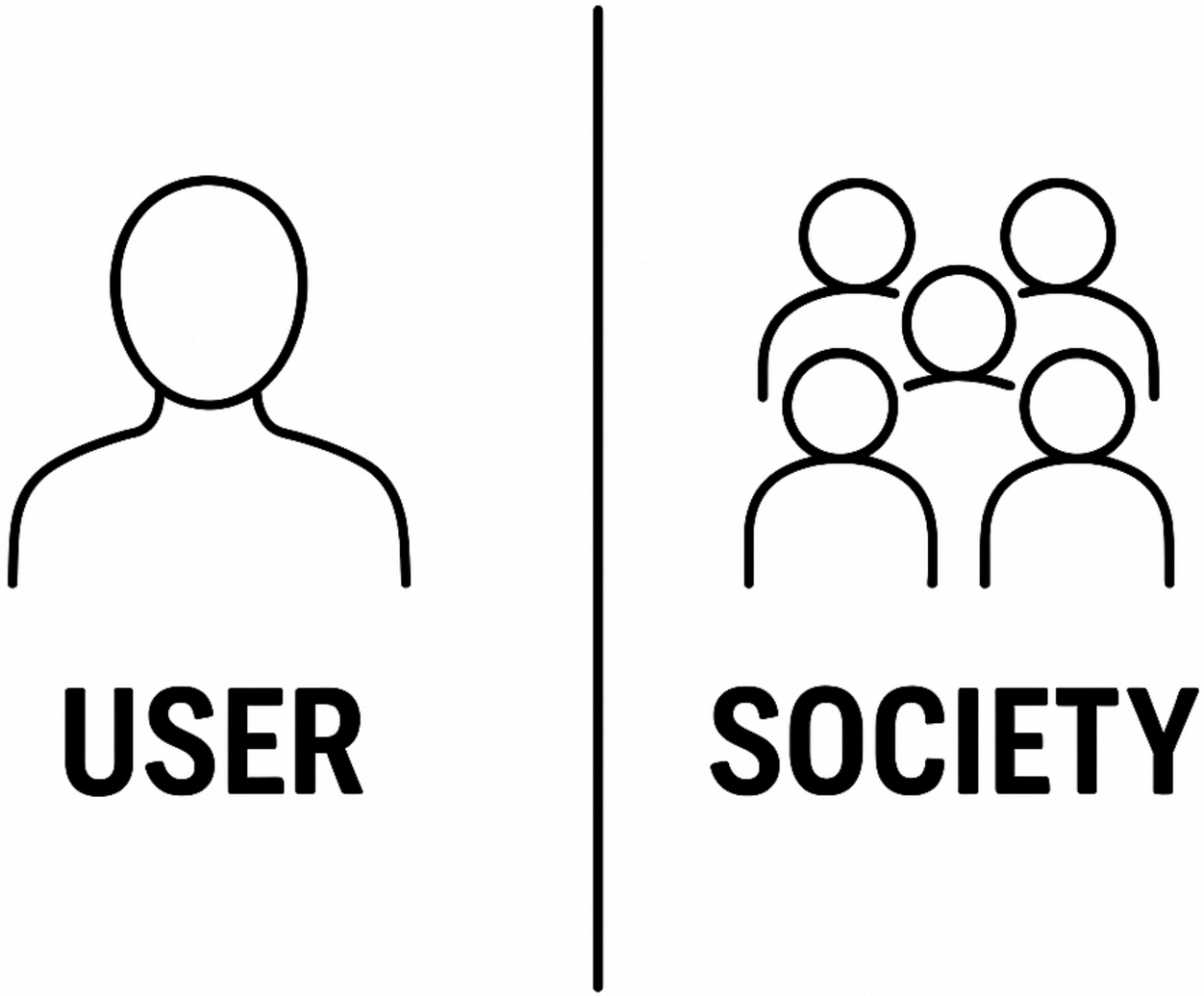
» If we hit **75%** by Monday, I will bring 🍪🍩🍫 at the presentations!

Responsible & Ethical Design; Trust & Transparency

When "Human-Centered" is Not Enough

We design systems for people.

But systems that benefit individuals can still harm society.



Why This Is Important in HCI

Designing for individuals is necessary — but no longer sufficient.

HCI must address societal, institutional, and ethical consequences of the systems we create.

The Readings

1. Six Grand Challenges for Human-Centered AI[1]
2. Perceptions of Algorithmic Decisions[2]
3. Overreliance & Cognitive Forcing[3]
4. Model Cards[4]

---

[1] Garibay et al. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *IJHCI*.

[2] Lee (2018). Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*.

[3] Buçinca et al. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *CSCW*.

[4] Mitchell et al. (2019). Model Cards for Model Reporting. *FAccT*.

Theme 1: What Does Human-Centered AI Require?[1]

[1] Garibay et al. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *IJHCI*.

Human-Centered AI

AI must be:

» Well-being oriented

» Responsible

» Privacy-sensitive

» Governed

» Cognitively compatible

A Real Question

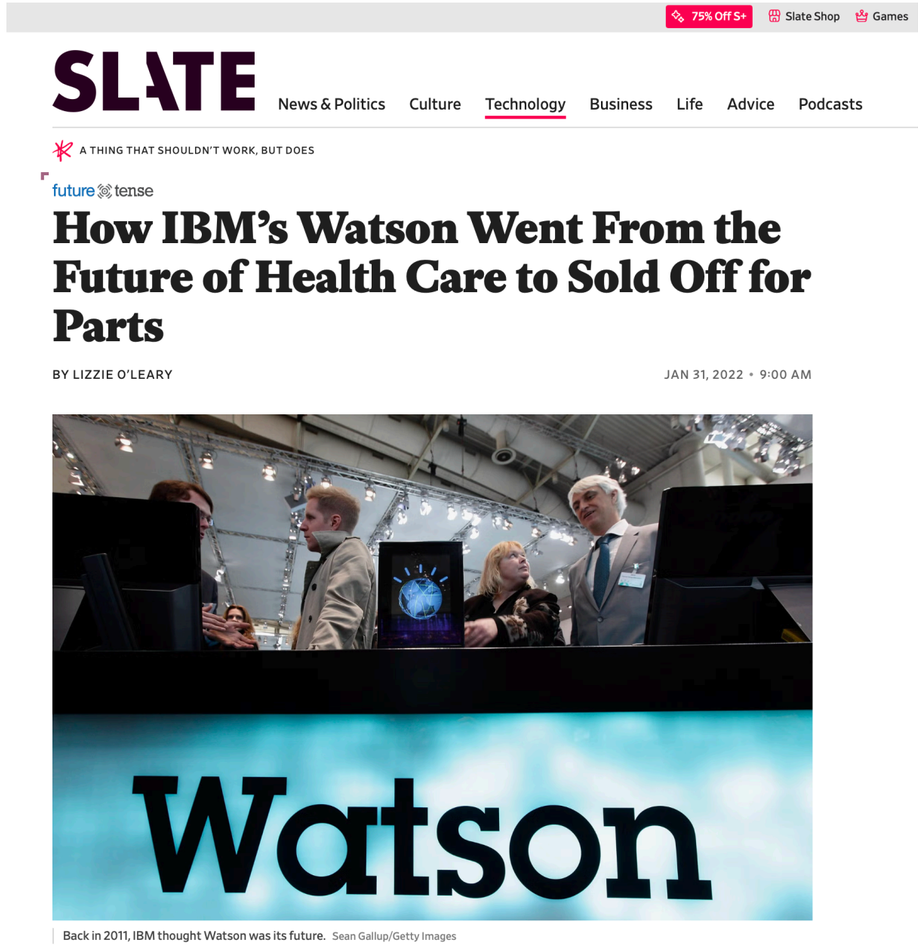
Why do some systems get trusted and adopted — while others get abandoned?

Let's look at some examples...

Example 1: IBM Watson for Oncology

Why abandoned:

- Doctors did not trust its recommendations.
- Recommendations were sometimes unsafe or clinically incorrect.
- It was not transparent about how decisions were made.
- Felt like a “black box” inserted into medical judgment.



75% Off S+ Slate Shop Games

SLATE News & Politics Culture Technology Business Life Advice Podcasts

A THING THAT SHOULDN'T WORK, BUT DOES

future tense

How IBM's Watson Went From the Future of Health Care to Sold Off for Parts

BY LIZZIE O'LEARY JAN 31, 2022 • 9:00 AM

Back in 2011, IBM thought Watson was its future. Sean Gallup/Getty Images

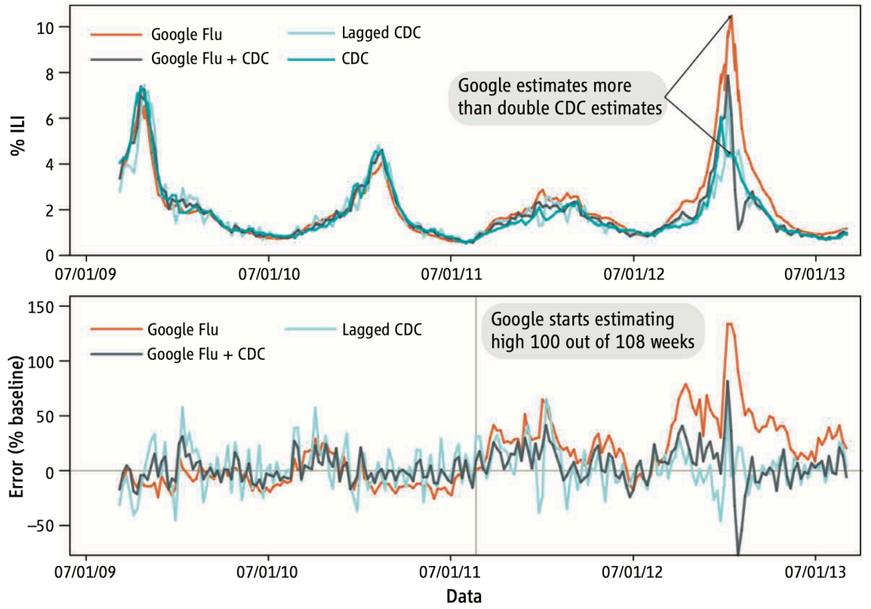
Example 2: Google Flu Trends[5]

Why abandoned:

» Initially seen as a breakthrough but became wildly inaccurate.

» Overfitted to search behavior shifts & media trends.

» Lost credibility with epidemiologists and the public.



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage [[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)]. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

[5] Lazer et al. (2014). The parable of Google Flu: traps in big data analysis. *science*, *343*(6176), 1203-1205.

Theme 2: Trust, Fairness & Emotion[2]

[2] Lee (2018). Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*.

Algorithmic Decisions Feel Social

People judge systems by:

» Fairness

» Process

» Respect

» Emotion

Two Failure Modes

1. Under-trust → avoidance, workarounds
2. Over-trust → automation bias, harm

Theme 3: Overreliance on AI[3]

[3] Buçinca et al. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *CSCW*.

Friction Helps People Think

Cognitive forcing functions:

» Ask for justification

» Reveal uncertainty

» Prompt alternatives

Outcome may be better decision-making.

Why Systems Fail

They collide with:

» Social conditions

» Inequalities

» Misaligned incentives

» Broken workflows

» Lack of recourse

Why Systems Succeed

They align with:

» Values

» Understanding

» Practice

» User agency

» Trust calibration

Theme 4: Model Cards[4]

[4] Mitchell et al. (2019). Model Cards for Model Reporting. *FAccT*.

Documentation as Intervention[6]

Model cards offer:

» Intended use

» Limitations

» Performance across groups

» Ethical considerations

[6] Model Card Toolkit

The Design Question

*What should responsible AI **feel** like?*

Patterns for Responsible AI

- Transparency
- Value alignment
- User agency
- Trust calibration
- Governance
- Cognitive respect

Looking Ahead

Leave with:

» Sensitivity to impact

» Critical awareness

» Tools for responsible design

» Vocabulary for trust

» Understanding of sociotechnical adoption

Closing

Design choices shape the world. What kind of world will our systems create?

Discussion

- We'll let AI randomly pick 3-5 names
- In the selected order, students:
  - Present their provocation/critical artifact/policy or design recommendation (30 secs)
  - Lead class discussion (5-8 min)