Human-Computer Interaction

Human-AI Interaction

Professor Bilge Mutlu

Today's Agenda

- Topic overview: *Mixed-Initiative, Complementarity & Evaluation*
- The frame: *enduring fundamental → how AI changes it*
- This week's readings
- Provocation: *complementarity or comfortable reliance?*
- Discussion

This term's frame

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

- **Fundamental:** the *division of initiative* — and the harder bar, does human+AI beat *either alone*?
- **How AI changes it:** fluent explanations inflate acceptance → **calibrated reliance**
- **Depth: Deep** — the keystone AI week

Topic overview: *Human-AI Interaction*

What do we mean by "AI"?

AI-infused systems: systems with features harnessing AI capabilities that are *directly exposed to the end user*.[1]

Not background automation – the human is *in the loop*, and the question is how initiative is divided.

[1] Amershi et al. (2019). Guidelines for Human-AI Interaction. *CHI 2019*.

Recap: the symbiosis vision[2]

Man-Computer Symbiosis, J.C.R. Licklider, 1960

"Men will set the goals, formulate the hypotheses… Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions…"

The unfinished promise: *who does what, and when?*

[2] Licklider, J.C.R. (1960). Man-Computer Symbiosis. *IRE Trans. Human Factors in Electronics*, HFE-1, 4–11.

This week's readings

- **Required**
  - Dell'Acqua et al., *The Jagged Technological Frontier* (2026)
  - Buçinca, Malaya & Gajos, *To Trust or to Think* (CSCW 2021)
- **Recommended** – Horvitz (1999) · Bansal et al. (2021)
- **Optional** – Amershi et al. (2019)

The anchor: mixed-initiative[3]

Each agent (human or computer) contributes **what it is best suited to, at the most appropriate time.**

Horvitz names the division of initiative — *who acts, when, and what the system does when it isn't sure what you want* — the direct ancestor of human-AI collaboration.

The question is when a system should take initiative, not just whether it can.

[3] Horvitz, E. (1999). Principles of Mixed-Initiative User Interfaces. *CHI 1999*.

Mixed-Initiative Levels[4]

Level	Capability
Unsolicited reporting	Agent notifies of critical information as it arises
Subdialogue initiation	Agent initiates dialogue to clarify, correct
Fixed subtask initiative	Agent takes initiative on predefined subtasks
Negotiated mixed initiative	Agents coordinate & negotiate to determine initiative

[4] Allen et al. (1999). Mixed-Initiative Interaction. *IEEE Intelligent Systems*.

Guidelines as consolidated practice[1]

1. **Initially** (G1–G2)
2. **During interaction** (G3–G6)
3. **When wrong** (G7–G11)
4. **Over time** (G12–G18)

Amershi's **18 guidelines** say how an interaction should *behave*. None says: **better than what?**

[1] Amershi et al. (2019). Guidelines for Human-AI Interaction. *CHI 2019*.

Why a human is in the loop at all

- Climb the ladder of automation and you lose sight of **where and why** it's wrong
- Monitoring a machine you can't predict is **its own workload**
- Shneiderman (Week 3): refuse the slider — **control *and* automation**[7]

👉 The human in the loop is this literature's **fix**. Today asks whether it works.

[7] Sheridan & Verplank (1978), tabulated in Parasuraman, Sheridan & Wickens (2000); discussed in Shneiderman (2020), the Week 3 reading.

10 levels of automation

Sheridan & Verplank (1978), as tabulated in Parasuraman, Sheridan & Wickens (2000)

	WHAT THE SYSTEM DOES	WHO HAS CONTROL
1	Human does everything	
2	Offers a full set of alternatives	
3	Narrows the set to a few	
4	Suggests one option	
5	Executes it — if the human approves	
	Initiative shift — it acts unless you stop it	
6	Executes unless the human vetoes in time	
7	Executes, then always informs	
	Transparency shift — it tells you only if you ask	
8	Executes, informs only if asked	
9	Executes, informs only if it decides to	
10	Decides and acts autonomously, ignoring the human	

Shaded band is our reading, not the paper's.

The harder bar: complementarity[5]

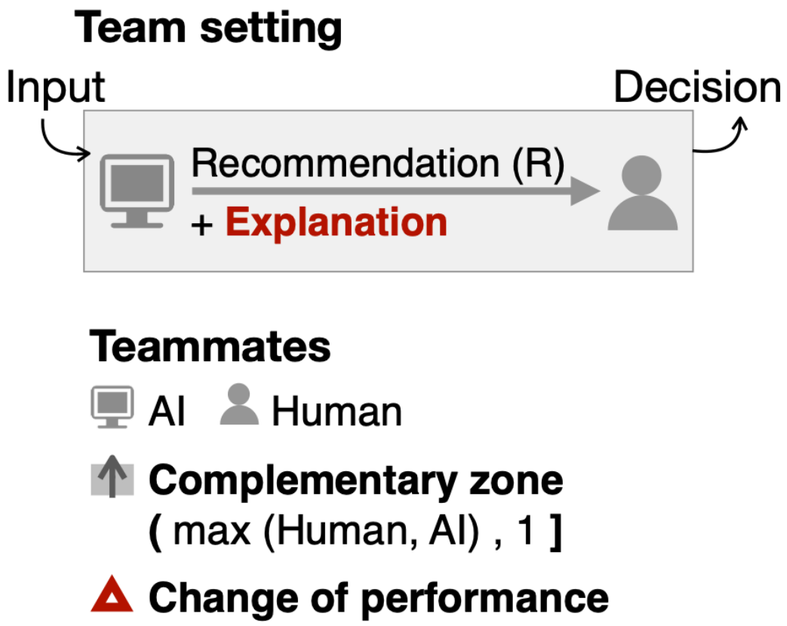
The fix has a premise: **the pair is better**. Better than *what*?

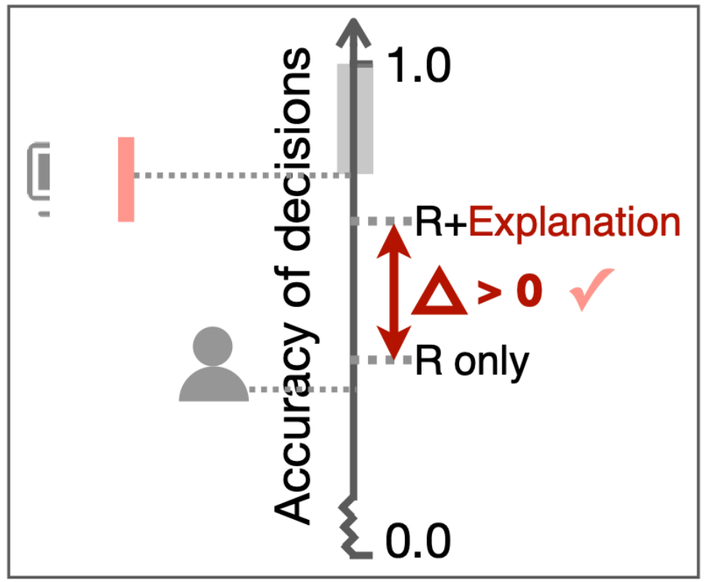
- **The bar:** the team has to beat **both** — the person working alone *and* the system working alone
- Earlier studies compared the team only to the **person**, in settings where the AI was already far more accurate

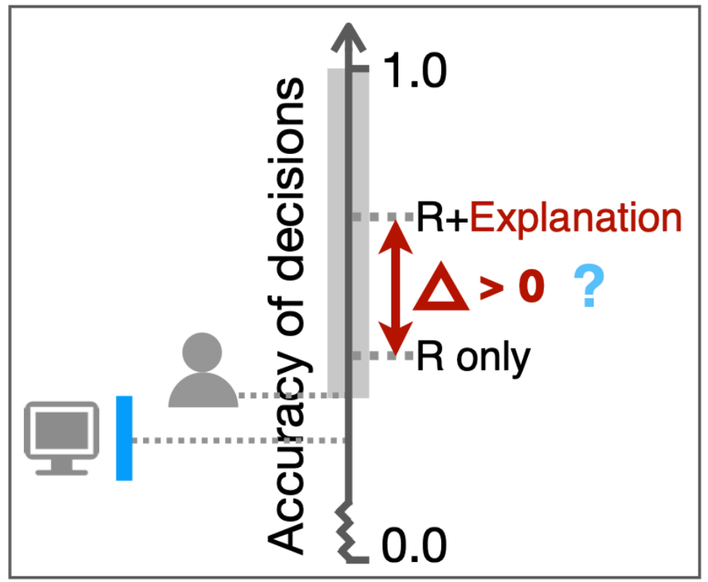
👉 If the team can't beat both, one of the two is **dead weight** — and that's a real decision someone has to make.

[5] Bansal, G. et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *CHI 2021*.

The complementary zone







So what happened?

1,626 people · 3 everyday tasks · AI tuned to human-level accuracy

- Teams **did** beat both solo arms — complementarity is achievable
- But explanations did **no better** than just showing the AI's confidence score
- And explanations made people **go along with the AI when it was wrong**

👉 Pairing helped. The *explaining* wasn't what helped.

The same bar, in the field[5b]

758 consultants. Real tasks. Randomized.

- **Inside** the frontier: quality **+30%**, time **−25%**
- **Outside** it: correctness **−19 percentage points**

The frontier is jagged

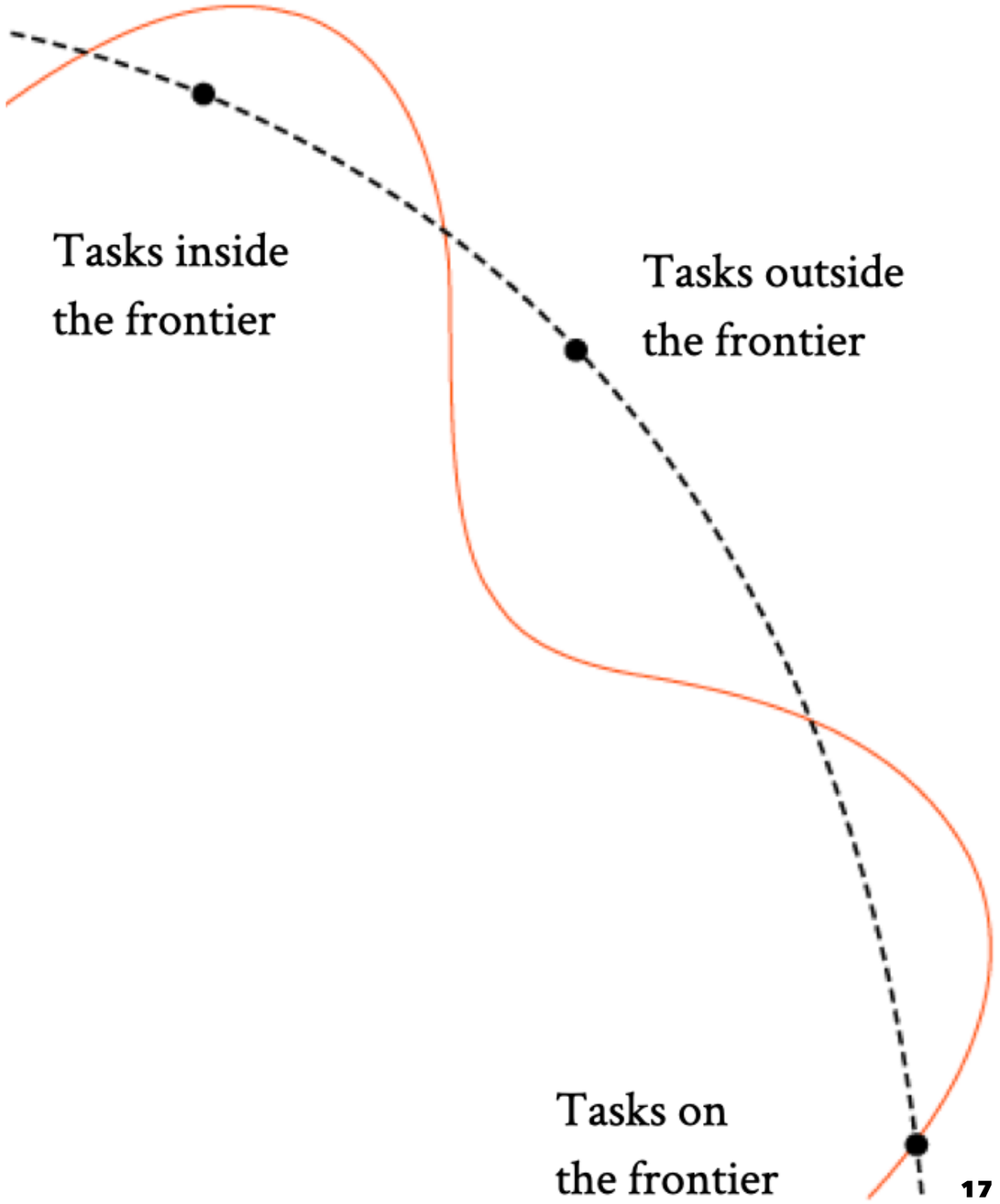
— capability is uneven, and the boundary is **invisible**.

[5b] Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz, H., Kellogg, K.C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K.R. (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*, 37(2), 403–423. doi:10.1287/orsc.2025.21838

Why "jagged"

The boundary isn't a tidy line from easy to hard.

- Neighbouring tasks land on **opposite sides**
- Nothing in the interface marks which side you're on



How would you *know*?

The only meaningful baseline is three-armed:

Human alone	AI alone	Human + AI
?	?	?

If human+AI underperforms a solo arm, **that is the finding.**

Calibrated vs. comfortable reliance[6]

We all do it: the suggestion appears, it looks reasonable, you take it.

Three ways to interrupt that reflex:

- **On demand** – the AI stays hidden until you ask for it
- **Update** – you decide first, *then* see the AI and may revise
- **Wait** – a 30-second delay before the AI appears

[6] Buçinca, Z., Malaya, M.B. & Gajos, K.Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *CSCW 2021*.

The catch

- Cognitive forcing **reduced overreliance** — the simple explanations didn't
- People gave those same designs the **least favorable ratings**
- The gains went **disproportionately** to people already high in **Need for Cognition**

The designs that work are the ones users like least — and they don't help everyone equally.

But what about *agentic* AI?

The dialectic, today

Fluent LLMs make acceptance *feel* free.

- **Horvitz:** initiative is the design variable – *when* should it act?
- **Bansal:** explanations raise acceptance, not performance
- **Dell'Acqua:** the frontier is jagged, and unmarked
- **Buçinca:** comfort crowds out oversight

👉 Complementarity is a claim to **test**, not assume.

Guardrails coda

- Over-reliance on fluent output — and trust ≠ appropriate reliance
- Buçinca's forcing functions worked by changing the **interaction**, not the user's intentions
- A guardrail that depends on your *willpower* isn't a guardrail

👉 **The design test:** does it live in the interface, or in my good intentions?

Run it on yourself

Three arms, on **your** work this week:

- Would I have done better **alone**?
- Would it have done better **without me**?

Almost nobody asks the second one.

👉 Week 1's policy, restated: did the **ceiling** go up, or did the **floor** drop?

The question today's studies don't ask

None of today's readings follows anyone over time.

- Work whose product is the **artifact** — the artifact is the point
- Work whose product is **you** — the artifact is a by-product

What does repeated assistance do to the person who has to supervise it *next* time?

👉 Decoration is what oversight **decays into** — an open question, not a finding.

Your provocation (*post one slide*)

Pick one:

1. **Provocation** — does a named AI-assist system deliver *complementary* performance, or *comfortable reliance*?
2. **Critical artifact** — a real system that trades calibrated reliance for acceptance; annotate it
3. **Design / policy** — one principle for keeping oversight *meaningful*, not nominal

Discussion

Today's roles

Reporter	Skeptic	Connector
Seat 1	**Seat 2**	**Seat 3**

Same in every pod. **Seats 4, 5 and 6** 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 — Methods: RtD + Wizard-of-Oz**
  - **Read:** Zimmerman & Forlizzi (2014) · Kelley (1984)
  - **If you can:** LFH **Ch. 10** (usability testing / think-aloud)
  - **Due 9:45 am:** Method Choice + Ethics Plan *(individual)*
  - Post 3-line project status
- **Next milestone:** RQ + literature review, due **Fri Oct 2**