Human-AI Interaction, Lecture 6

Designing Human-AI Systems

Humane Launches Ai Pin - Marking A New Beginning for Personal AI Devices

hu.ma.ne

NEWS PROVIDED BY
Humane Inc. →
Nov 09, 2023, 12:00 ET

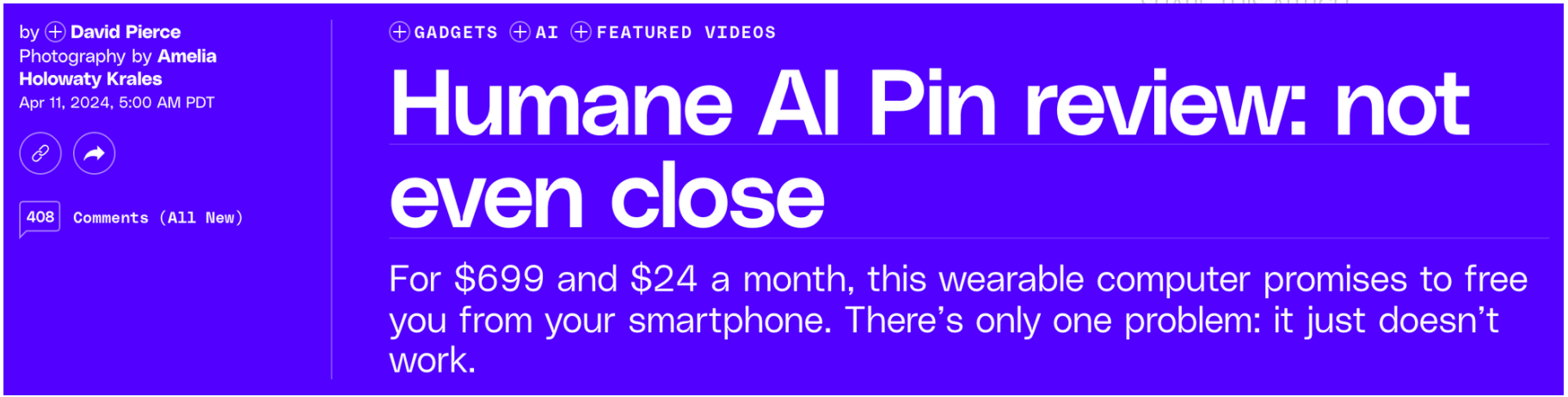
SHARE THIS ARTICLE

The Ai Pin projects text onto your palm instead of a screen.

Humane, via PR Newswire, Nov 9, 2023

Humane Launches Ai Pin - Marking A New Beginning for Personal AI Devices

hu.ma.ne



Humane, via PR Newswire, Nov 9, 2023; D. Pierce, "Humane AI Pin review: not even close," The Verge, Apr 11, 2024 (Verge score: 4 out of 10)

NEWS AI GADGETS

Humane is shutting down the AI Pin and selling its remnants to HP



Photo by Amelia Holowaty Krales / The Verge

/ HP is acquiring the Humane platform and team for $116 million.

by Jay Peters
Feb 18, 2025, 2:29 PM PST

178 Comments (All New)

If you buy something from a link, The Verge may earn a commission. See our ethics statement.

On February 28, 2025, the pins stopped connecting to Humane's servers.

J. Peters, The Verge, Feb 18, 2025; HP press release, Feb 18, 2025; Humane, Ai Pin Consumers FAQ

How do you design a product when you **can't be sure what the AI will do**?

Today

1. Why AI is hard to design
2. Start from people
3. Prototype the AI's behavior
4. Amershi's guidelines
5. Discussion
6. Guideline audit
7. When the AI is wrong

Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design

Qian Yang **Aaron Steinfeld** **Carolyn Rosé** **John Zimmerman**
Carnegie Mellon University, Pittsburgh PA, USA
{yangqian,steinfeld}@cmu.edu cp3a@andrew.cmu.edu johnz@cs.cmu.edu

Designers handle complex technology all the time. **What makes AI different?**

Two sources of difficulty

Capability uncertainty: “uncertainties surrounding what the system can do and how well it performs”

Output complexity: “complexity of the outputs that the system might generate”

Yang et al., CHI 2020

Output complexity

Write a short toast for my sister's wedding.

Run 1: one toast

Here is a short, warm, and versatile toast you can adapt. ...

“Good evening, everyone. For those who don’t know me, I’m [Your Name], [Sister’s Name]’s [older/younger] sibling. ...

Run 2: a toast, then tips

Here is a short, heartfelt, and versatile toast you can adapt for your sister’s wedding. ...

Tips for Delivery: ...
Stay sober (or nearly so): ...

Run 3: three options

Here are three short toast options, ranging from sweet to slightly humorous. ...

“... I remembered that [Sister’s Name] and I **used to share a bed when we were kids**, and ...”

The same request gives different answers.
The third run made up a childhood memory.

Qwen3.8-27B

Capability uncertainty

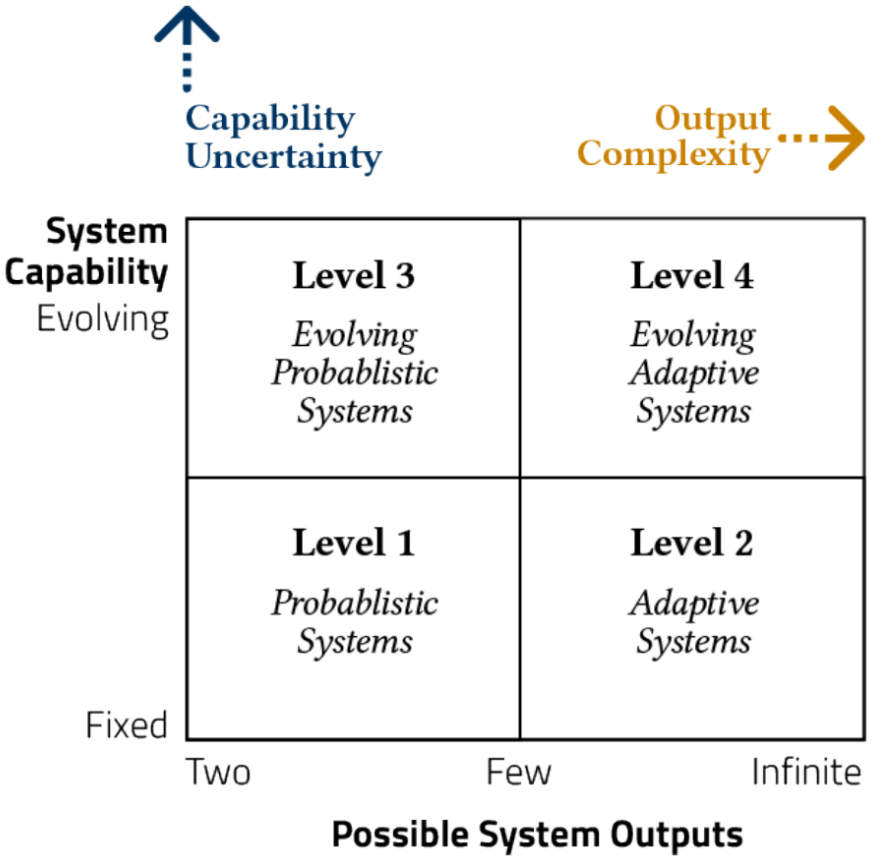
“It performs worse for less frequent users and less typical situations.”

Yang et al., CHI 2020

The pin: “too many things it does well but **only sometimes**”

Pierce, The Verge, Apr 11, 2024

Four levels of AI systems

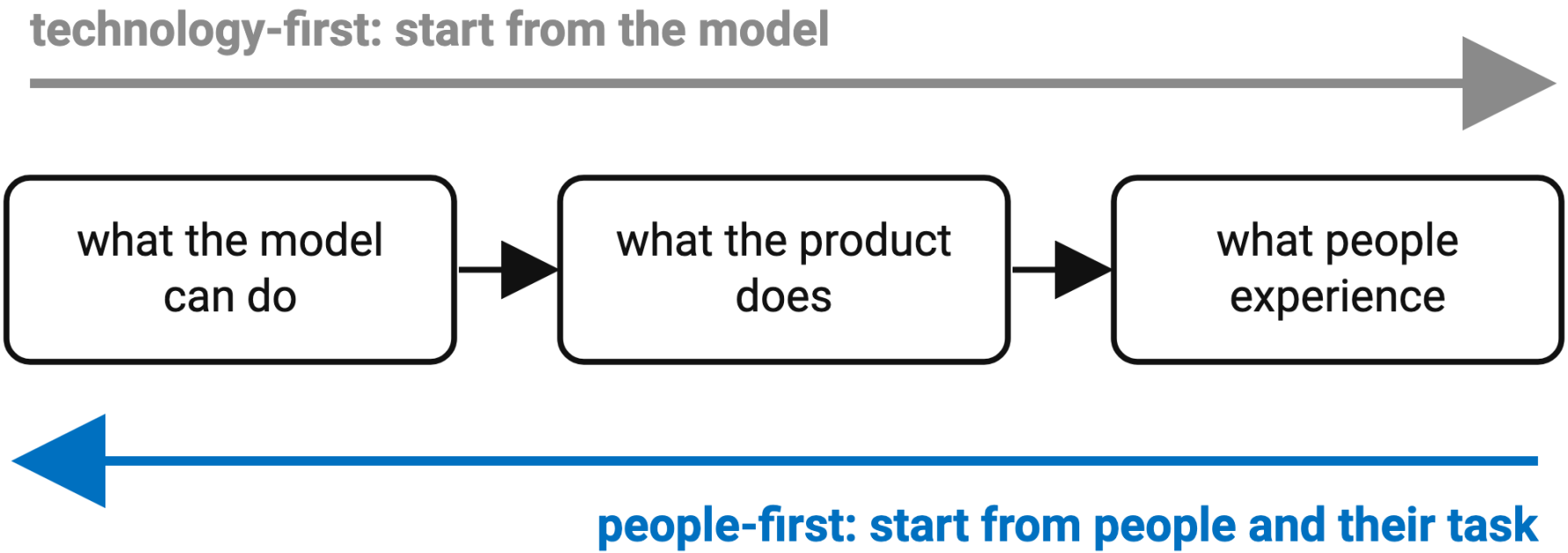


Level 1: usual design methods work

Level 4: capability changes after launch and outputs are open-ended

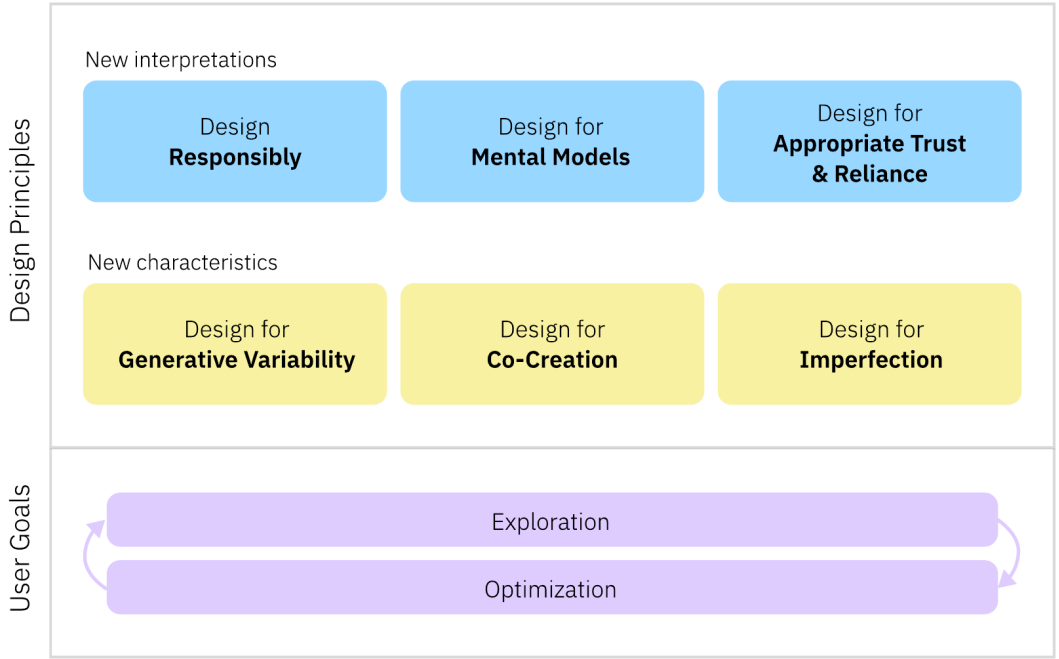
Yang et al., CHI 2020, Figure 4

Technology-first and people-first



Google PAIR, People + AI Guidebook (2025 edition); Yang et al., CHI 2020

“Design for the user by understanding their needs and pain points, and not for the technology or its capabilities.”



Weisz et al., CHI 2024, Figure 1

Does AI add value here?

CONSIDERATIONS

- Stakes →
- User context →
- Cost →

Stakes

Situations can have high or low stakes. Imagine the types of situations where people may use your system's outputs, and assess how much they should trust it to be accurate and precise, or creative and comprehensive.

Some parts of a task get **worse** with AI.

Google PAIR, People + AI Guidebook (2025 edition), “Decide if AI adds unique value”

Map the task

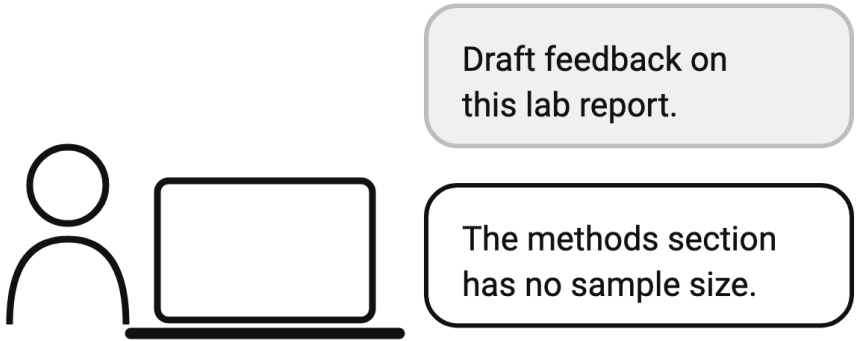
A TA writes feedback on lab reports



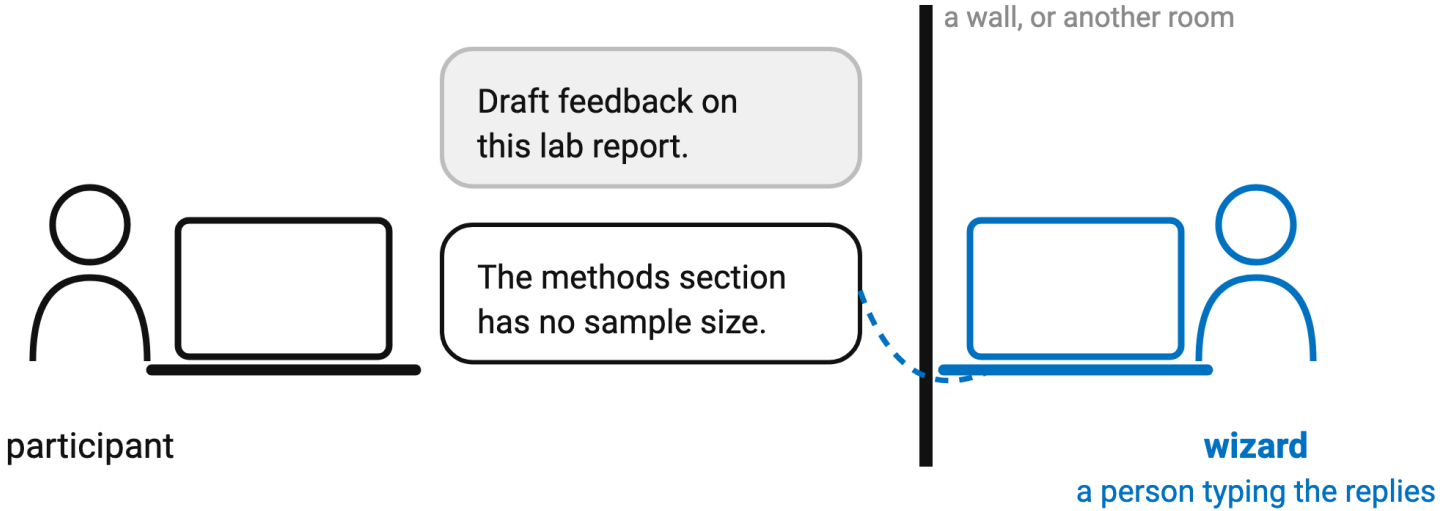
It's hard to sketch AI outputs **ahead of time**

Yang et al., CHI 2020

Wizard of Oz: a person plays the AI



Wizard of Oz: a person plays the AI

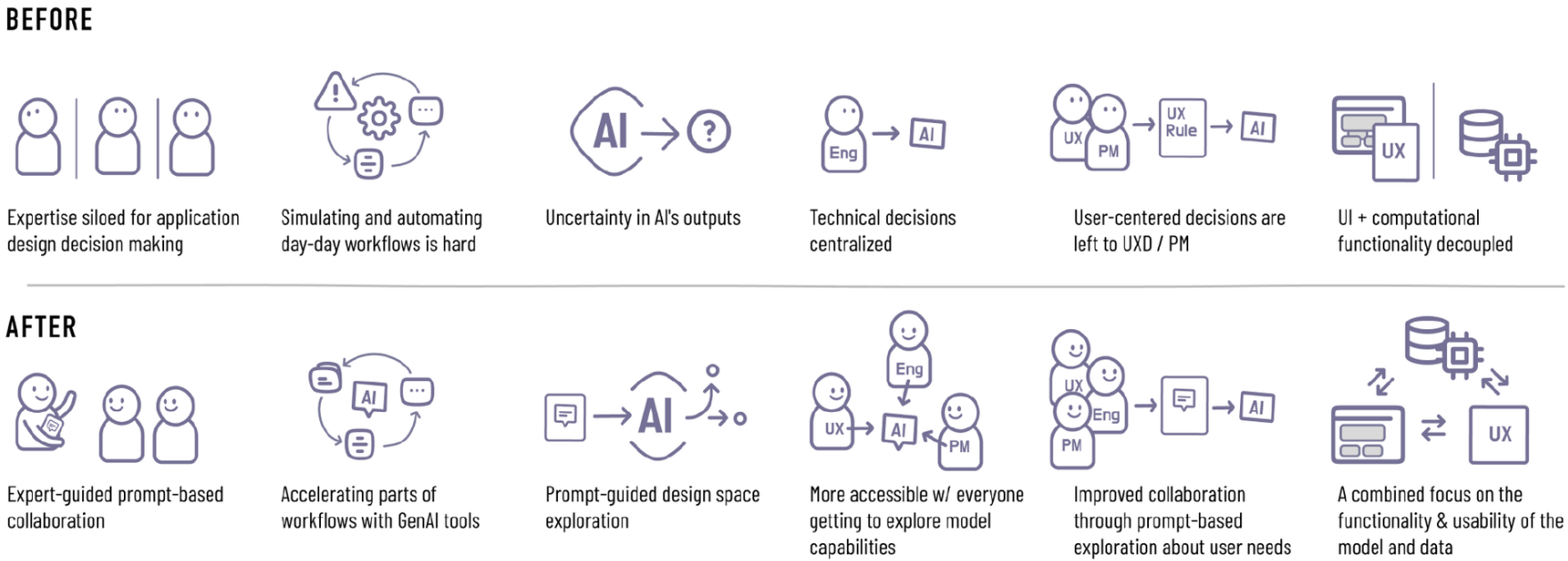


+ cheap and fast

- **it misses the AI's real errors**

Yang et al., CHI 2020

Prototyping with prompts



39 practitioners, 13 sessions. In each, a designer, an engineer and a product manager prototyped with prompts.

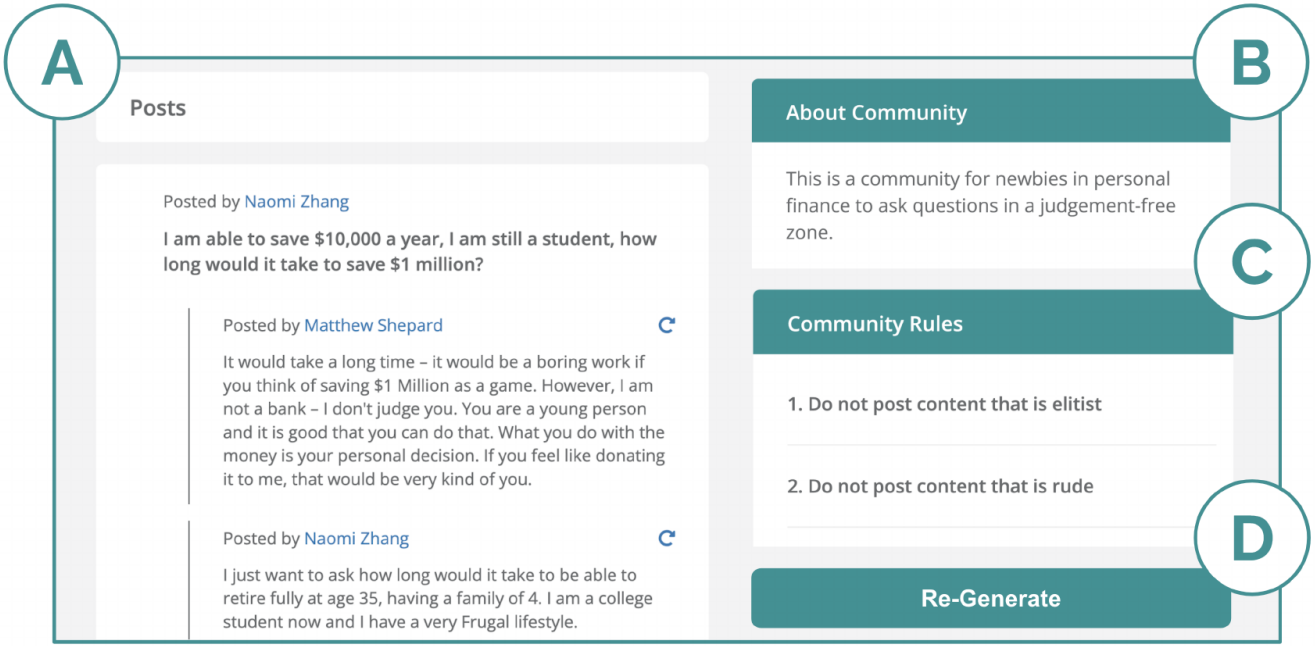
Subramonyam, Thakkar, Ku, Dieber & Sinha, CHI 2025, Figure 1 (before and after generative AI)

Prototyping with prompts: **a few examples**

“these examples are all very similar to the first example since they’re all generated from the first example”

a product manager in the study, Subramonyam et al., CHI 2025

Social simulacra



Shown one real and one generated thread, people picked the wrong one in 41% of pairs.

Park et al., UIST 2022, Figure 1: a generated community in SimReddit

Amershi et al.: 18 guidelines

Grouped by when they apply:

Initially, **During interaction**, **When wrong**, **Over time**

Amershi et al., CHI 2019

Examples from their study

Initially: G2, Make clear how well the system can do what it can do.

[Music Recommenders, Product #1] “A little bit of hedging language: ‘we think you’ll like’.”

During interaction: G4, Show contextually relevant information.

[Web Search, Product #2] “Searching a movie title returns show times in near my location for today’s date”

Amershi et al., CHI 2019, Table 1

Examples from their study

When wrong: G10, Scope services when in doubt.

[Autocomplete, Product #1] “It usually provides 3-4 suggestions instead of directly auto completing it for you”

Over time: G16, Convey the consequences of user actions.

[Social Networks, Product #2] “[The product] communicates that hiding an Ad will adjust the relevance of future ads.”

Amershi et al., CHI 2019, Table 1

Guideline 11: **Make clear why the system did what it did**

It had one of the most violations.

Amershi et al., CHI 2019

Discussion: Yang and Amershi

1. Where do a chat model and the AI Pin go on Yang's map?
2. Which of the 18 guidelines matter more for chat models? Less?
3. Which method would have caught the pin's problems early?

Guideline audit

Groups of 4, 15 minutes

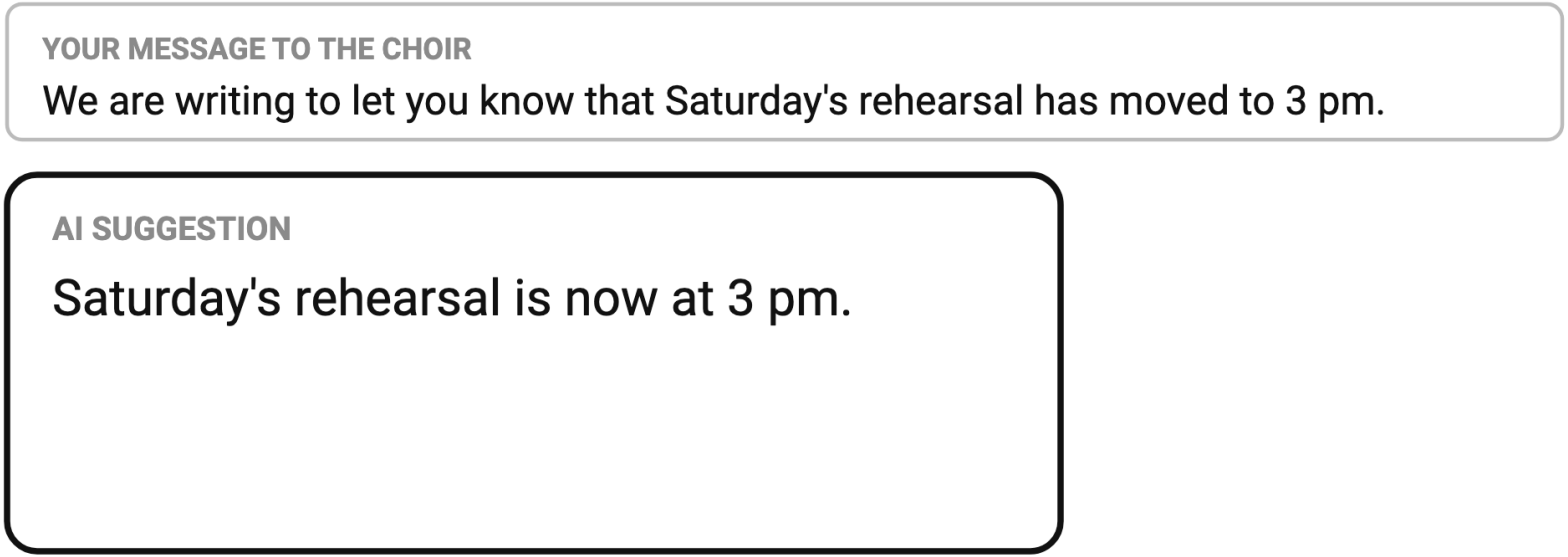
1. Count off 1 to 4: your number is your phase
2. Pick an AI product everyone has used
3. Find one guideline it breaks, and screenshot it

The 18 guidelines, by phase

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

Amershi et al., CHI 2019, Table 1, which gives each guideline's title, a description and an example

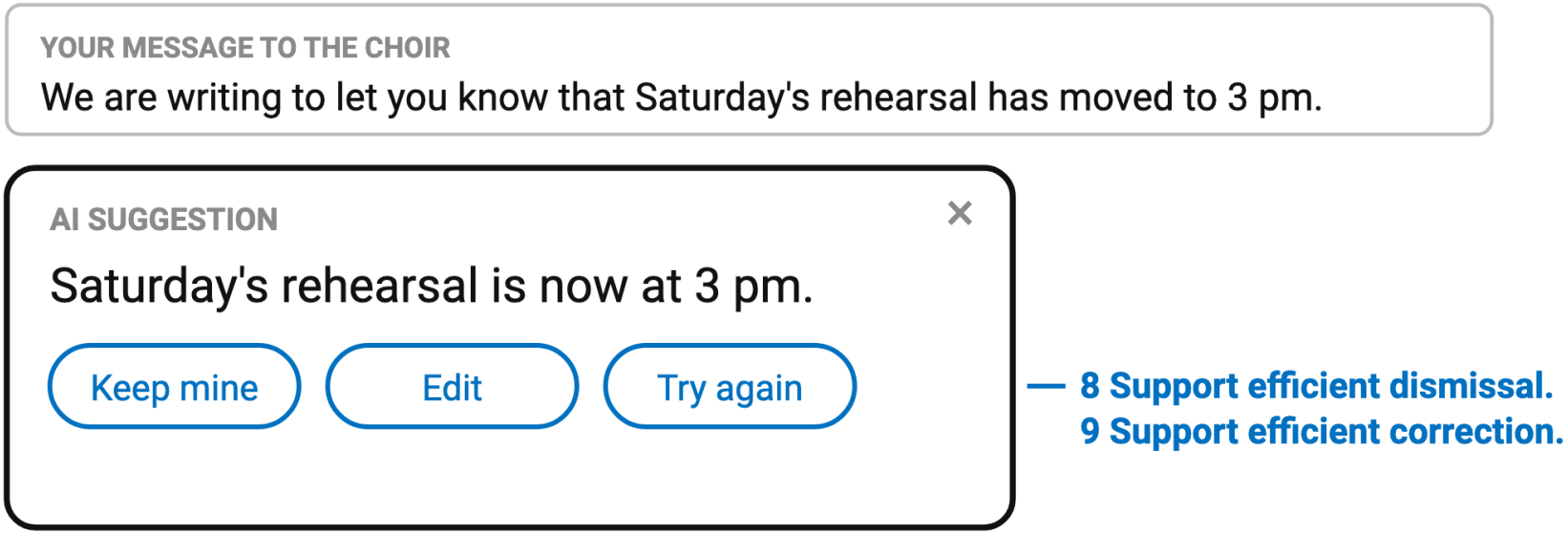
Designing for errors



“It’s worth assuming that risks and errors can’t be wholly prevented” (Google PAIR)

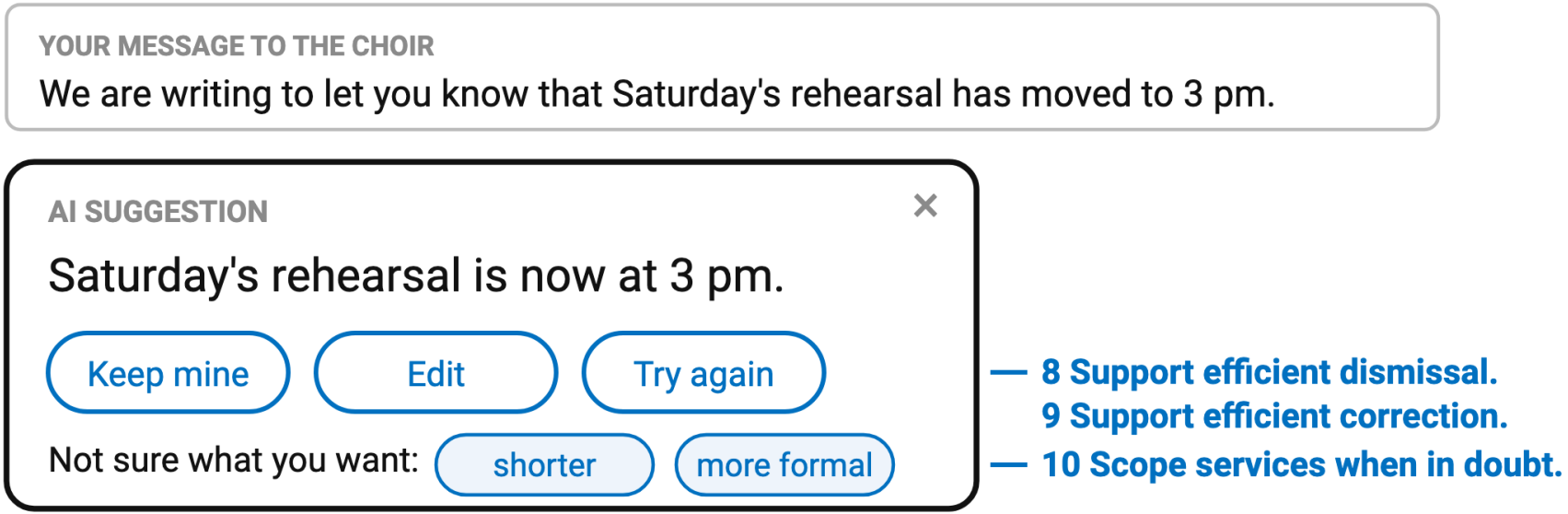
Suggestion text: Qwen3.8-27B

Designing for errors



Suggestion text: Qwen3.8-27B

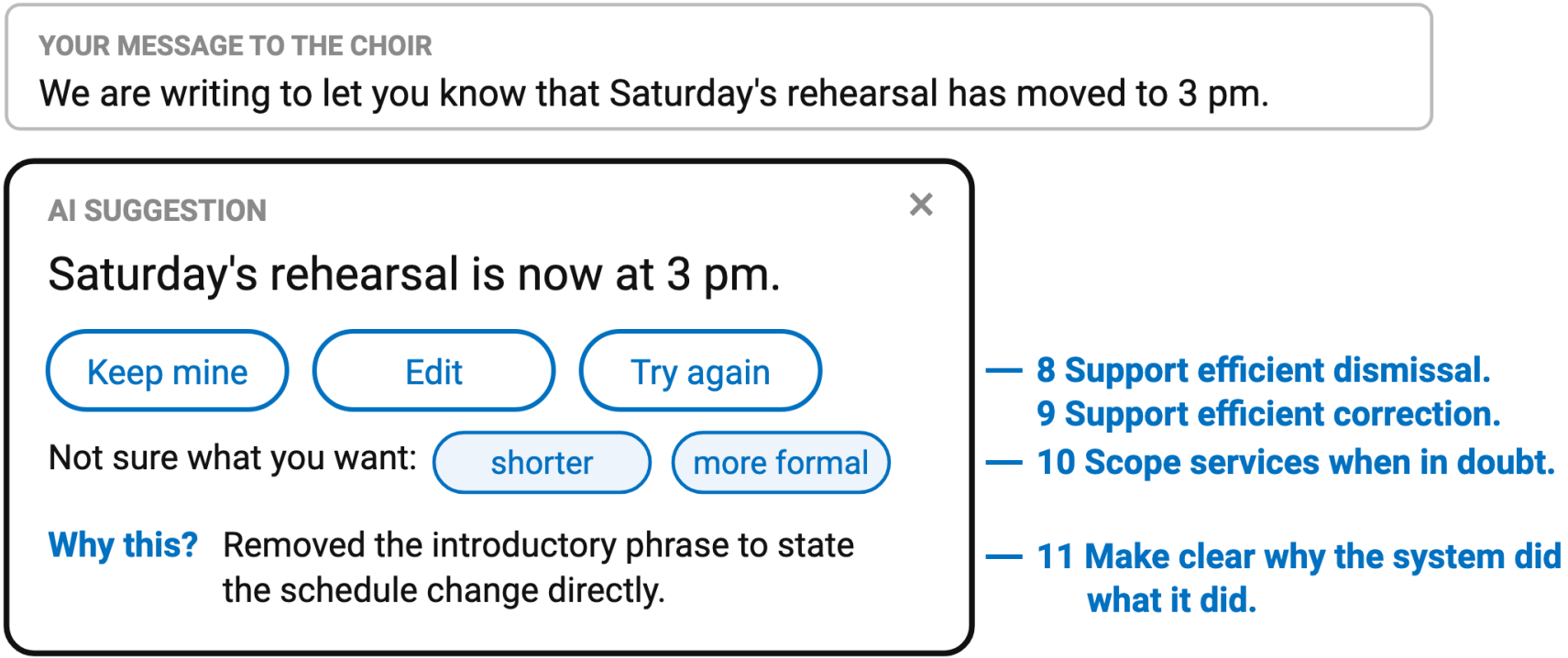
Designing for errors



— **8 Support efficient dismissal.**
9 Support efficient correction.
— **10 Scope services when in doubt.**

Suggestion text: Qwen3.8-27B

Designing for errors



— **8 Support efficient dismissal.**
9 Support efficient correction.

— **10 Scope services when in doubt.**

— **11 Make clear why the system did what it did.**

Suggestion and reason: Qwen3.8-27B

The Ai Pin and guideline 10

- Me: “Play ‘Texas Hold ’Em’ by Beyoncé.”
- The AI Pin: “Songs not found for request: Play Texas Hold ’Em by Beyonc\u00e9. Try again using your actions find a relevant track, album, artist, or playlist; Create a new PlayMusic action with at least one of the slots filled in. If you find a relevant track or album play it, avoid asking for clarification or what they want to hear.”

Guideline 10: when unsure, **ask the person**.

D. Pierce, The Verge, Apr 11, 2024

Milestone 2: Make it understandable



Milestone 1	**Milestone 2**	Proposal	Milestone 3	Milestone 4	Demo fair
due today	due Thu, week 5	Tue, week 6	Thu, week 8	Tue, week 10	finals

Due Thursday of week 5

Before Thursday

for Lecture 8, Uncertainty and Explanation

1. Write your tool's prompt
2. Run real requests several times each
3. Mark the wrong outputs

Takeaways

1. What an AI can do is **uncertain**, and its outputs **vary**.
2. Start from the people and their **task** before picking a model.
3. Prototype with a **real model** to find its errors early.
4. Plan for **wrong outputs**, and make them easy to **fix**.

Next: Lecture 7

Evaluating Human-AI Systems

Read before class:

Bansal et al., CHI 2021

Lee et al., TMLR 2023, Sections 1-2 (skim Section 3)

Vaccaro et al., Nature Human Behaviour 2024