Methods in AI Research 11
Designing responsible AI part 2:
Designing & Evaluating AI and Automated systems

Chris Janssen

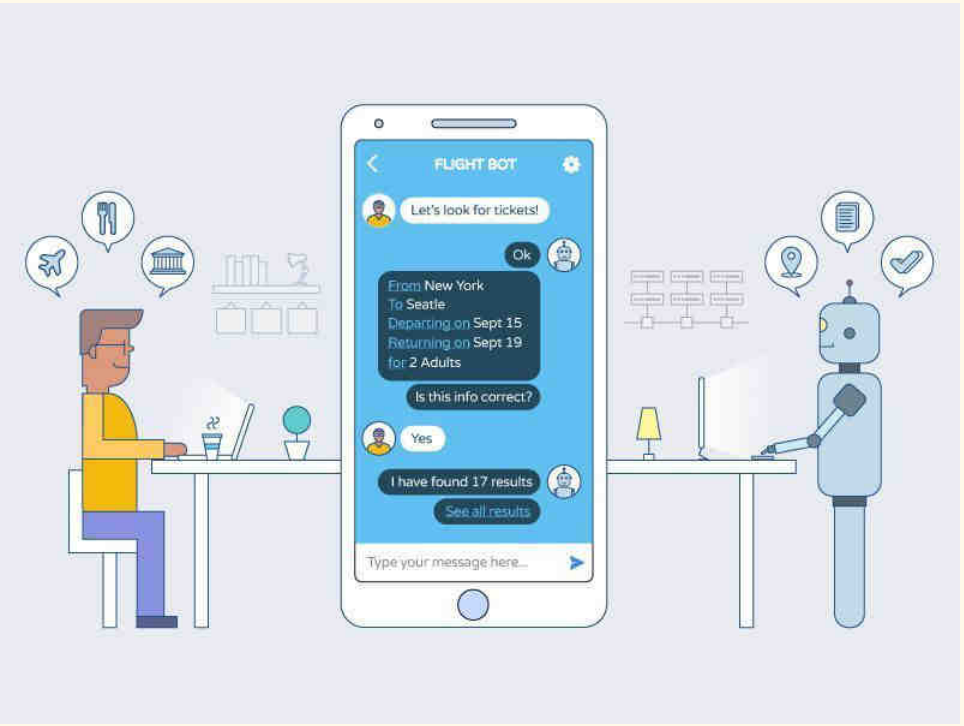
c.p.janssen@uu.nl
www.cpjanssen.nl

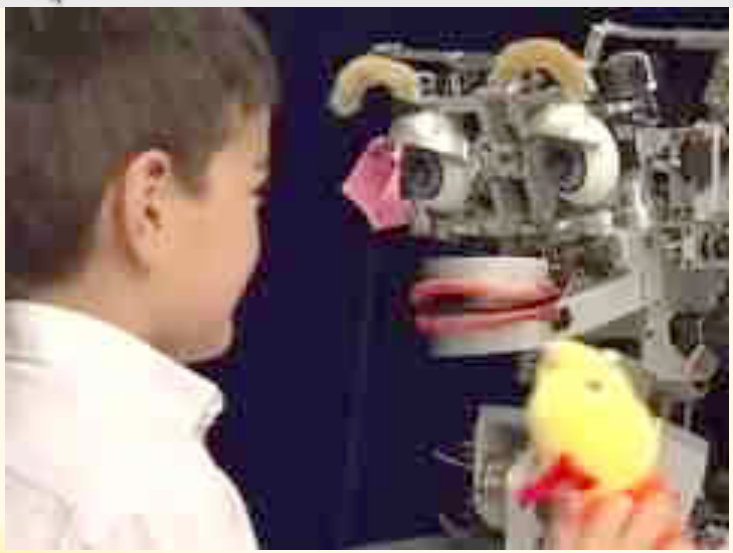
Exam – Not part of it

Why consider design?

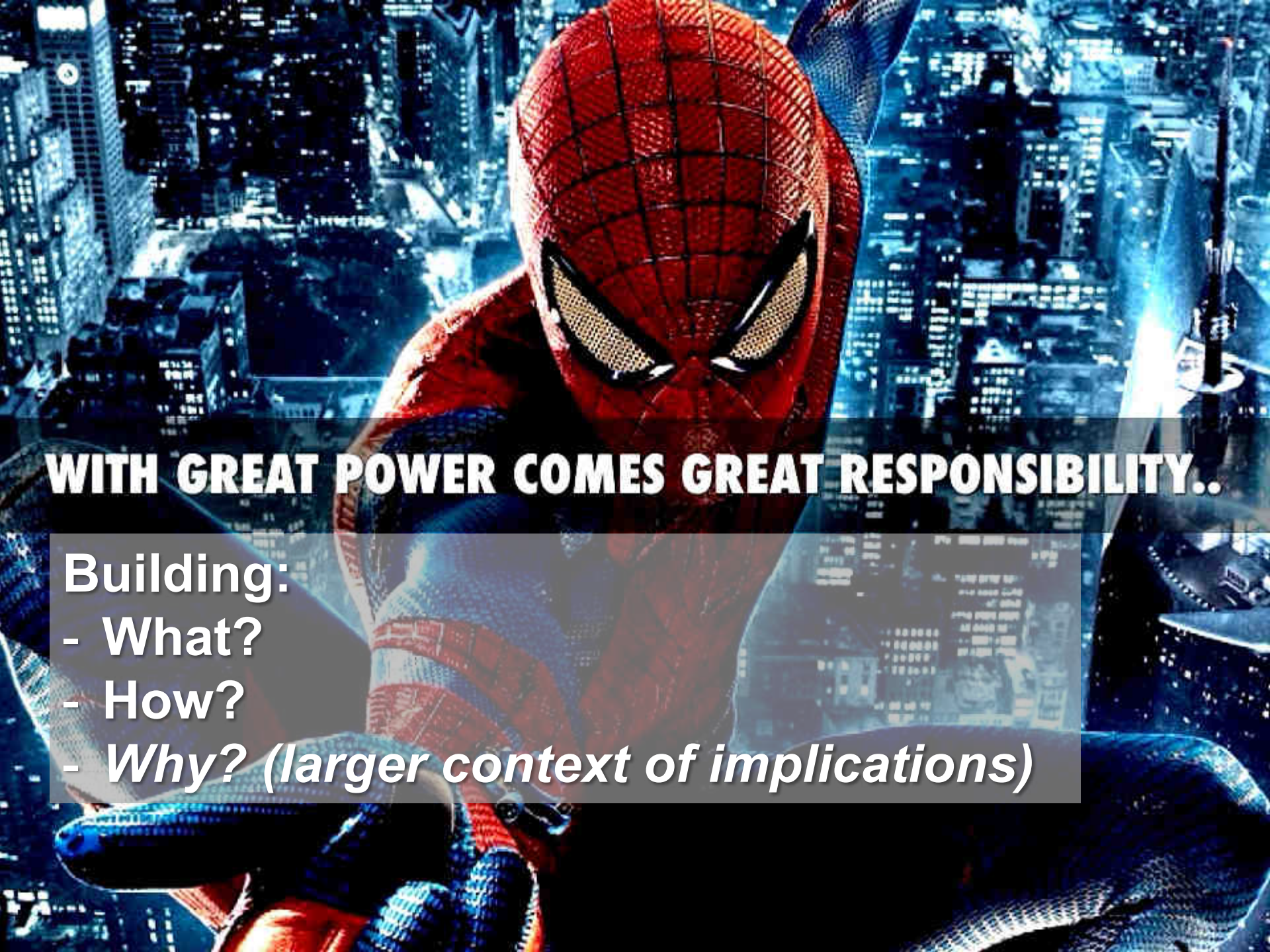
AI: understanding by building

DESIGN





facebook



Building:

- **What?**
- **How?**
- ***Why? (larger context of implications)***

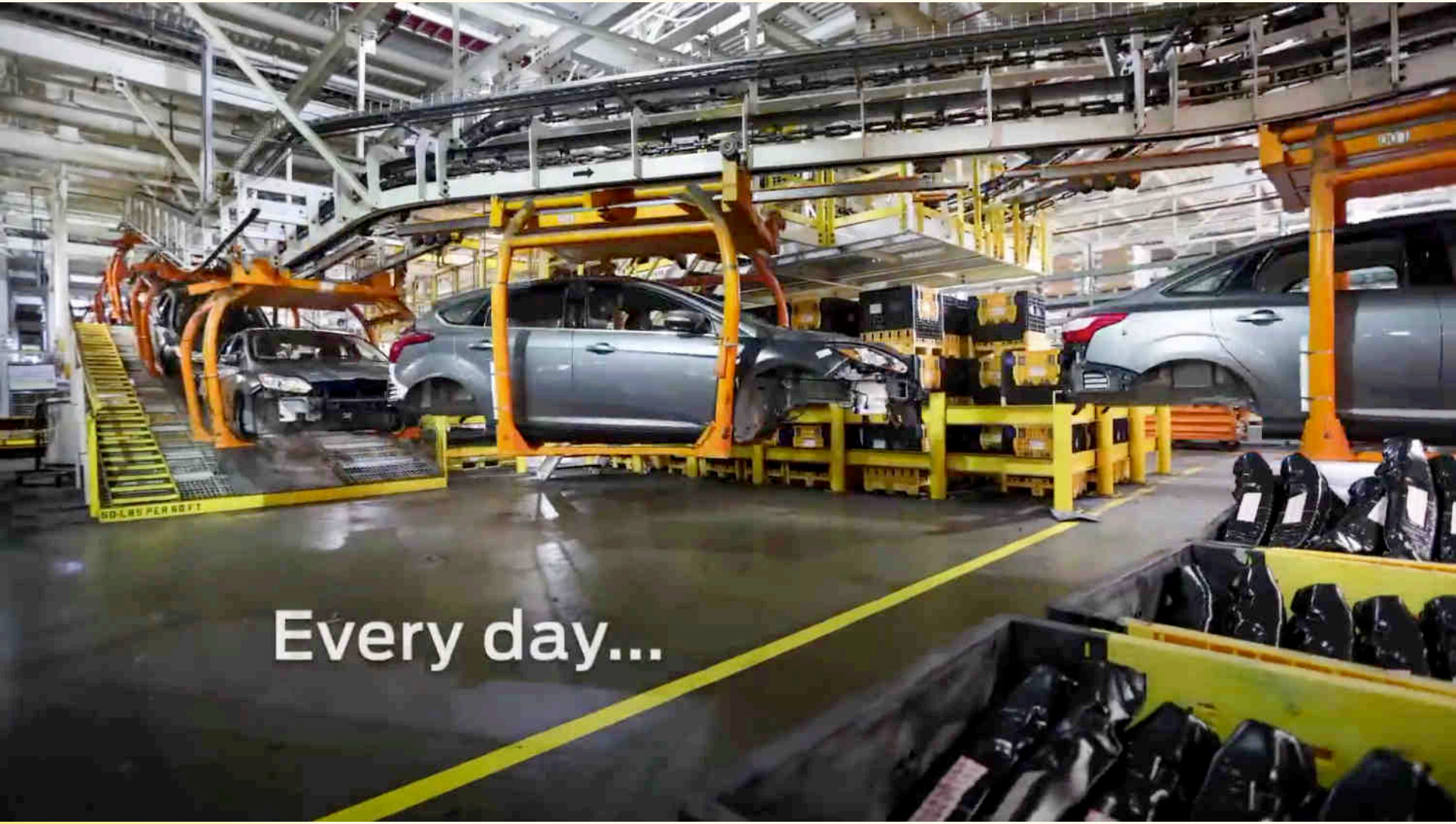
Today's topics

- Why consider design at all?
  - AI: Understanding by building
- Brief history of human-automation (and AI) interaction
- Radical / disruptive innovations
- Why is designing for human-automation (and AI) interaction tricky?
  - Dynamic humans, systems, contexts, interaction
- Guidelines for design
  - Microsoft's guidelines for Human-AI Interaction
  - Lee, Wickens, Liu & Boyle's human-centric guidelines

- If you want to know more..

Examples of automation?

Start (according to some): Ford's factory assembly lines

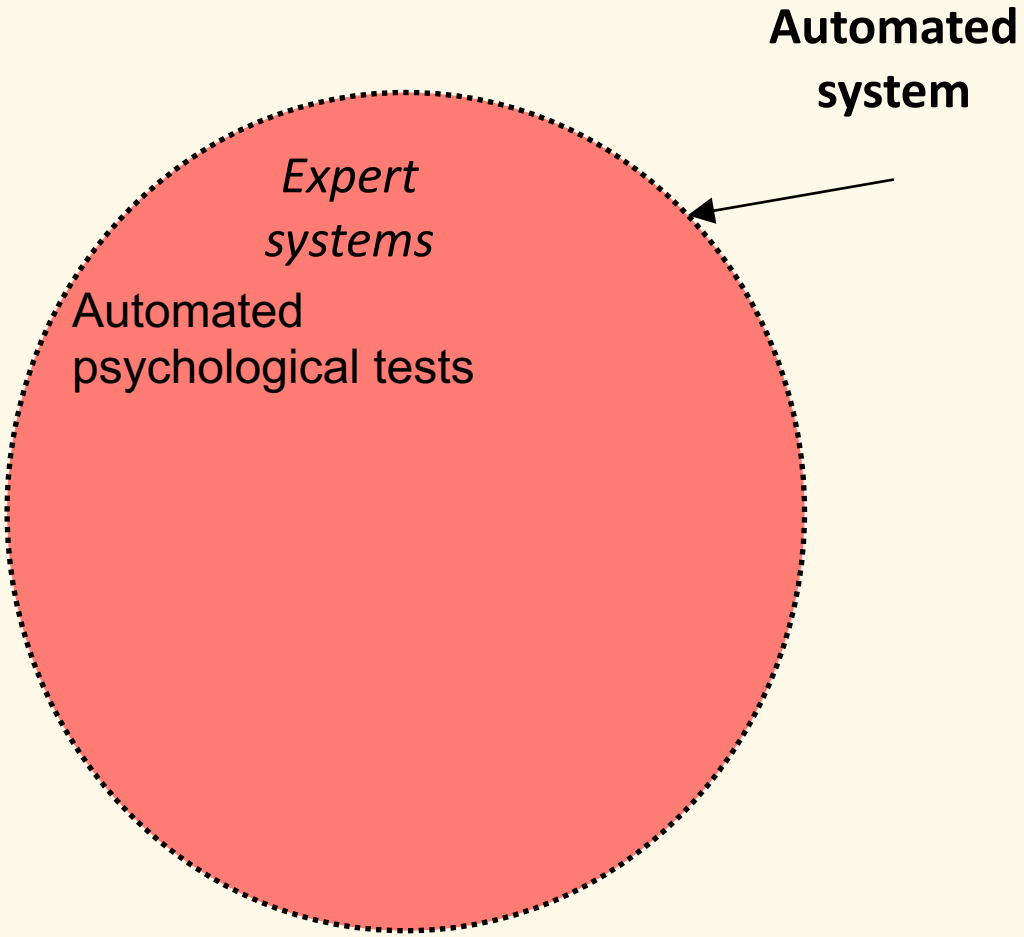


https://www.youtube.com/watch?v=jLud5XYfY_c

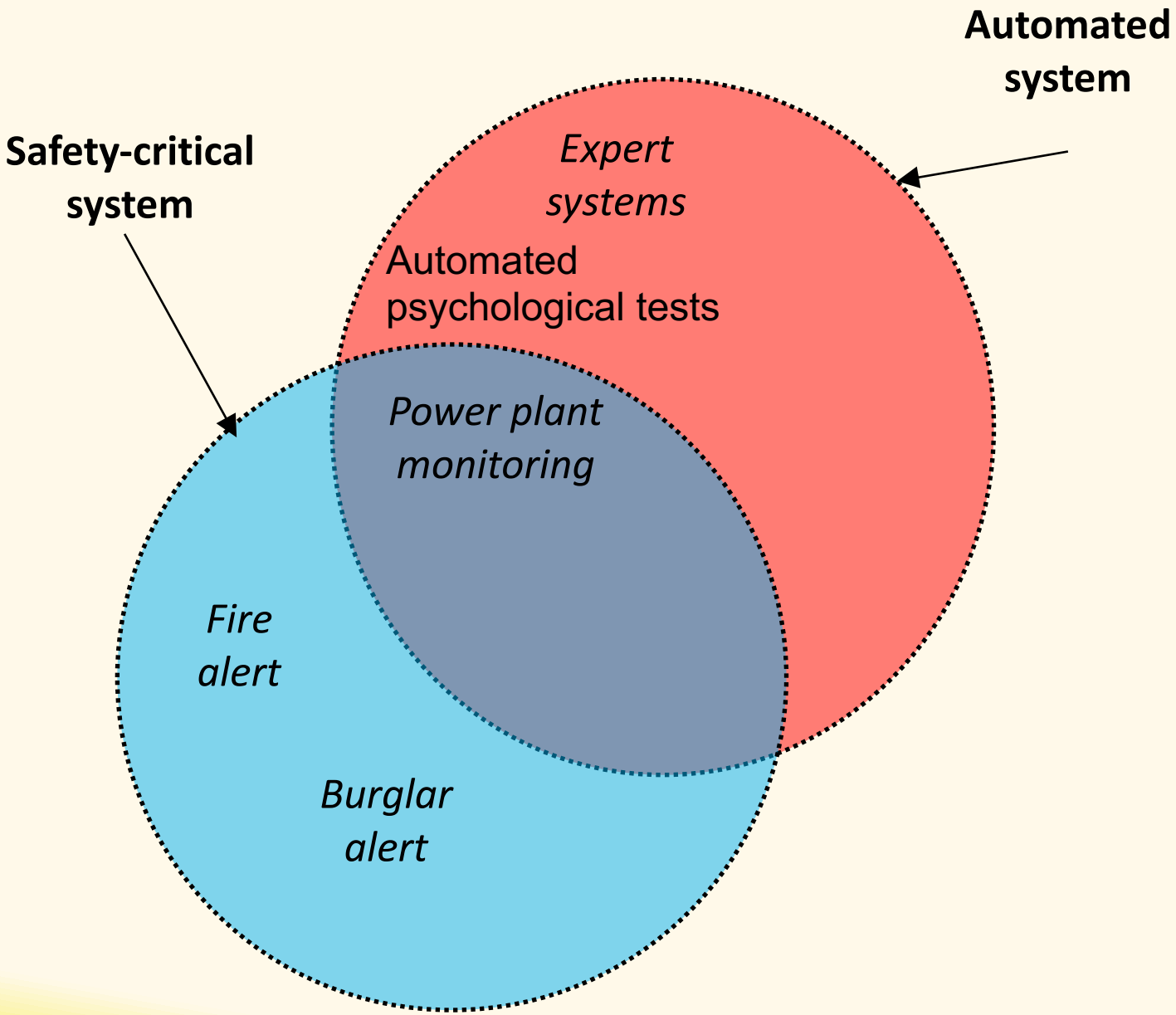
Automation (Brittannica encyclopedia)

"*the **application of machines** to **tasks once performed by human** beings or, increasingly, to **tasks** that **would otherwise be impossible**. Although the term mechanization is often used to refer to the simple replacement of human labour by machines, **automation generally implies the integration of machines into a self-governing system**.*"

- No mention of computer processor; no hard requirement
- "Self-governing system" -> AI

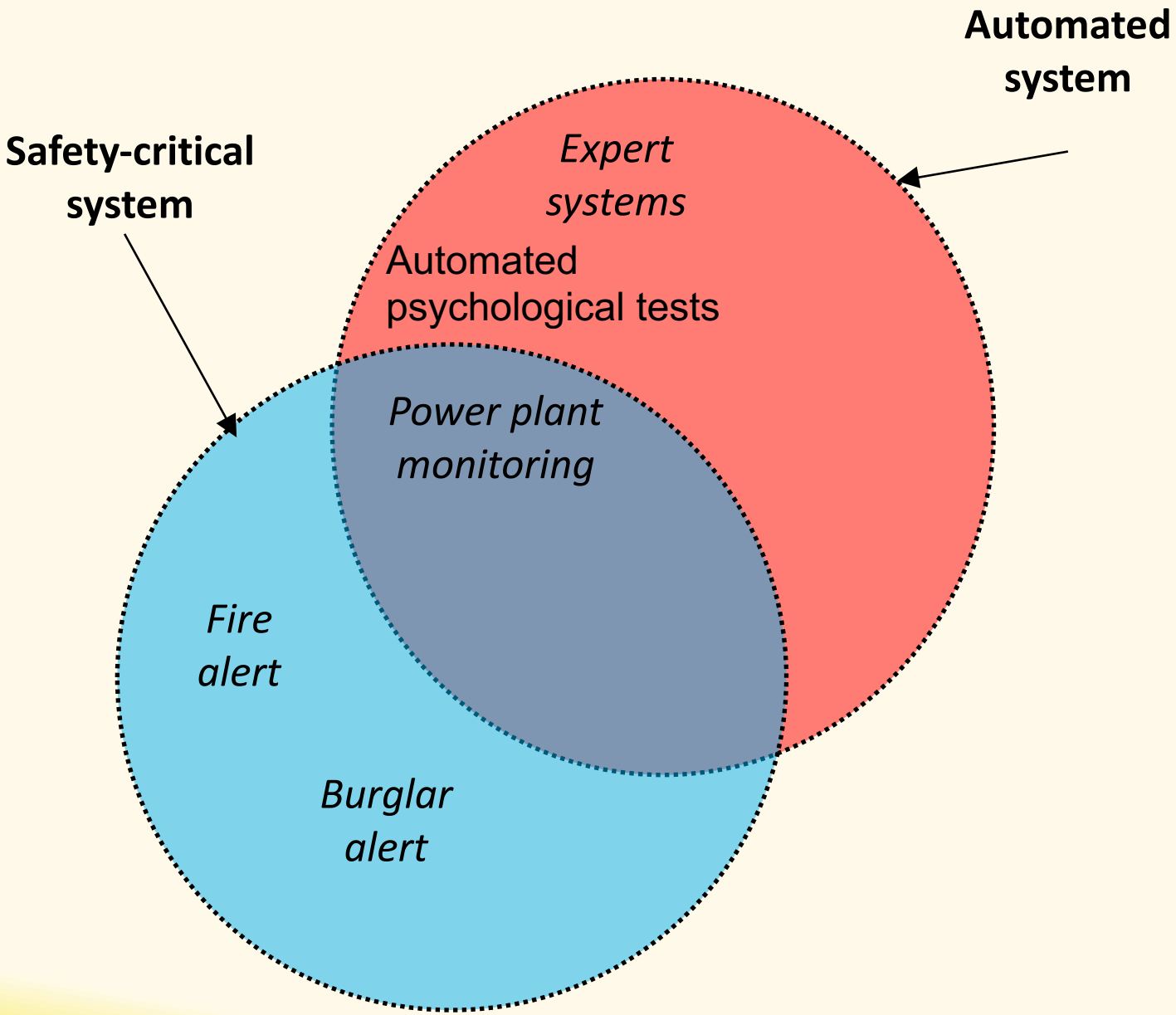


Janssen, Donker, Brumby, Kun (2019)

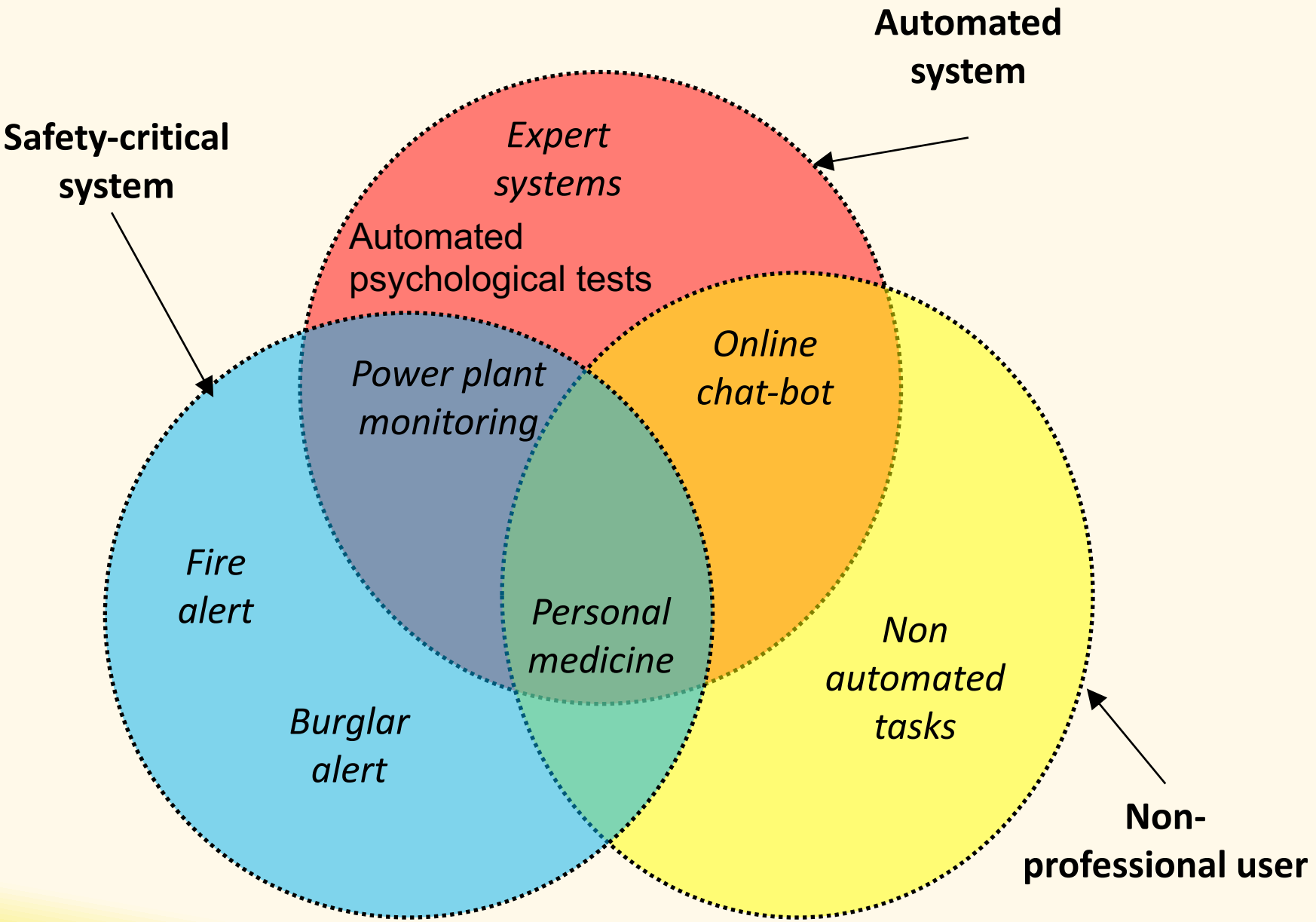


Janssen, Donker, Brumby, Kun (2019)

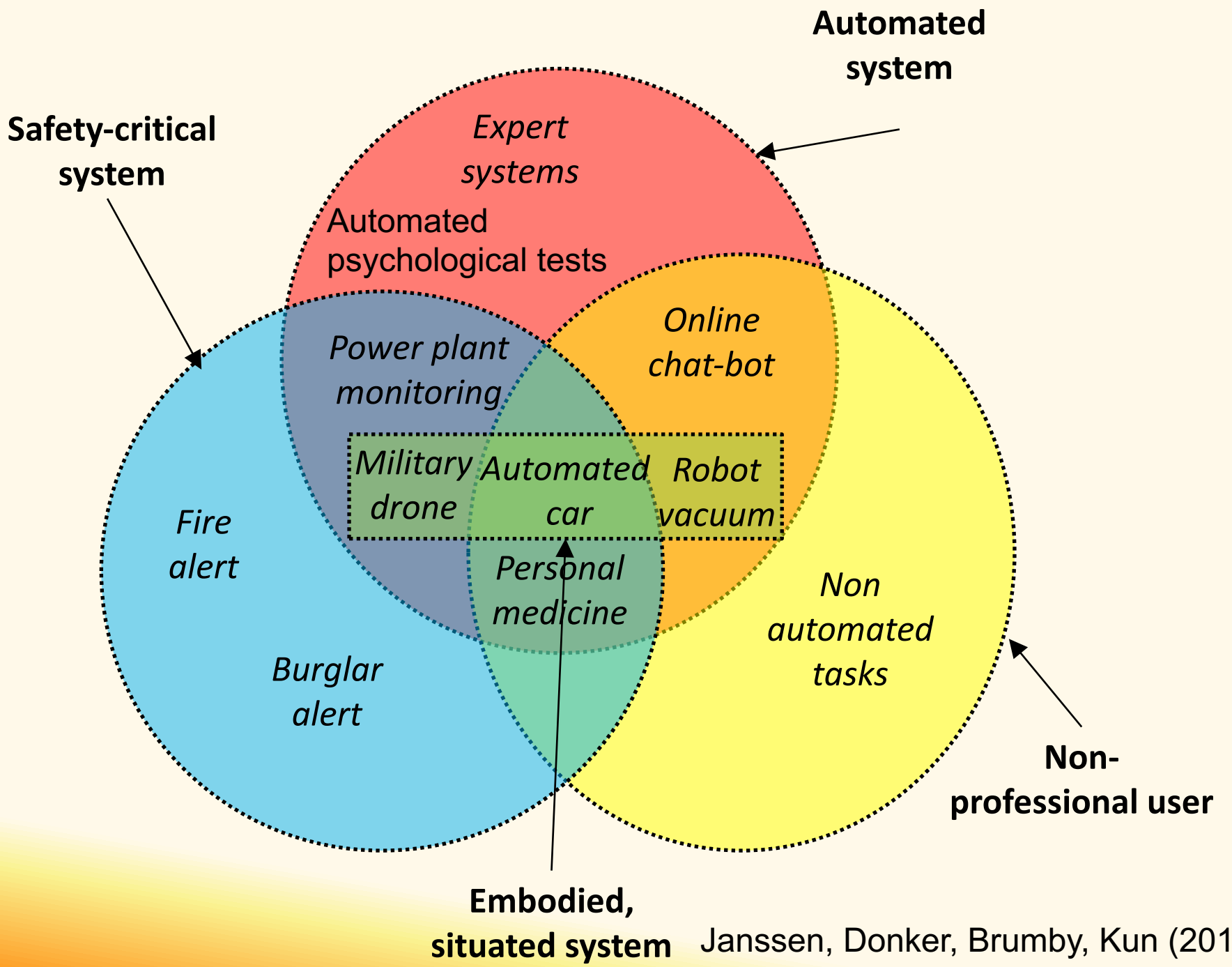
http://gpti.in/images/upcoming_programs/controlsystem.png



Janssen, Donker, Brumby, Kun (2019)



Janssen, Donker, Brumby, Kun (2019)



Janssen, Donker, Brumby, Kun (2019)

History & future of human-automation interaction

- **More breadth:**
  - **Safety-critical systems**
  - **Things that "act in the world": embodied, situated systems (which is a noisy environment!)**
  - **Non-professional users: no training**

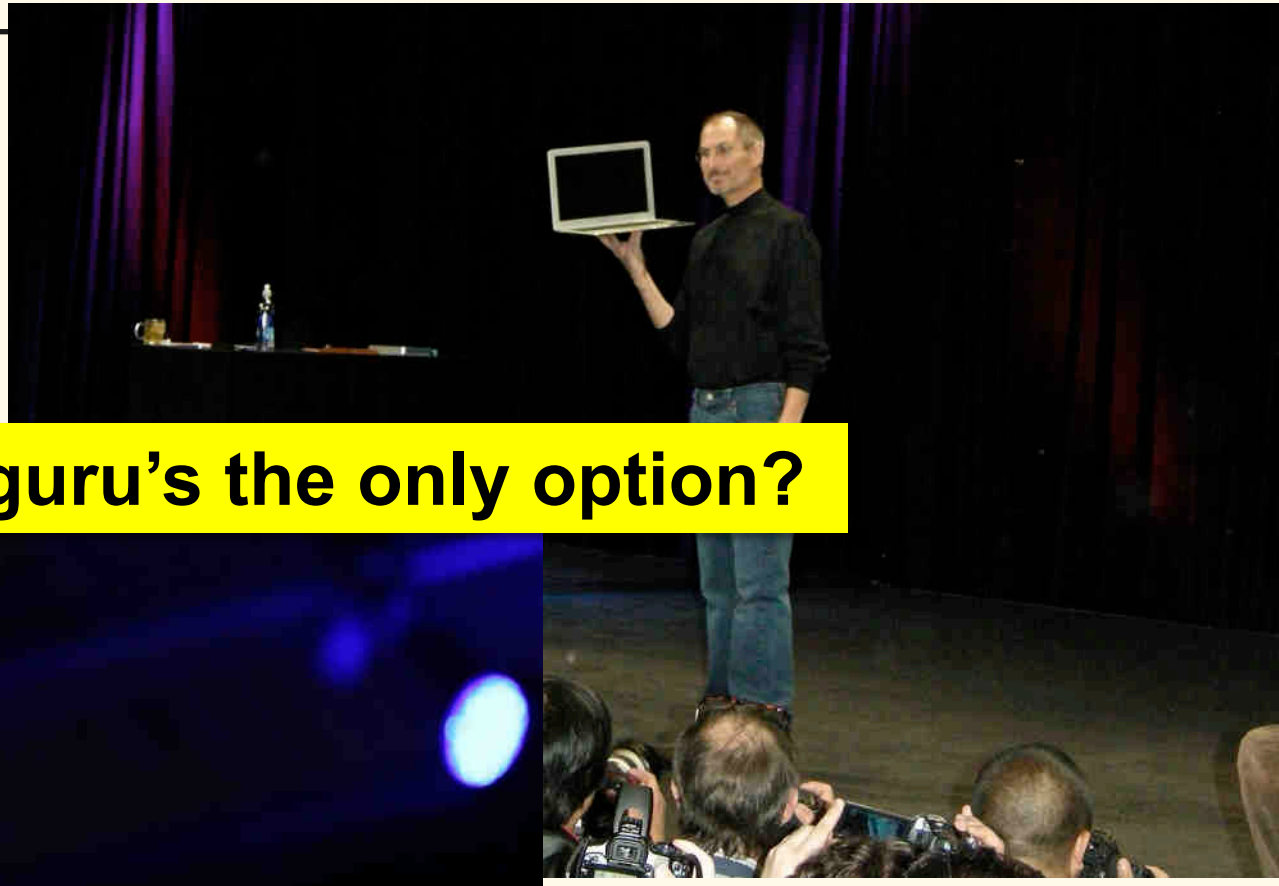
- **More depth needed: how to "design for all"**

Janssen, Donker, Brumby, Kun (2019)

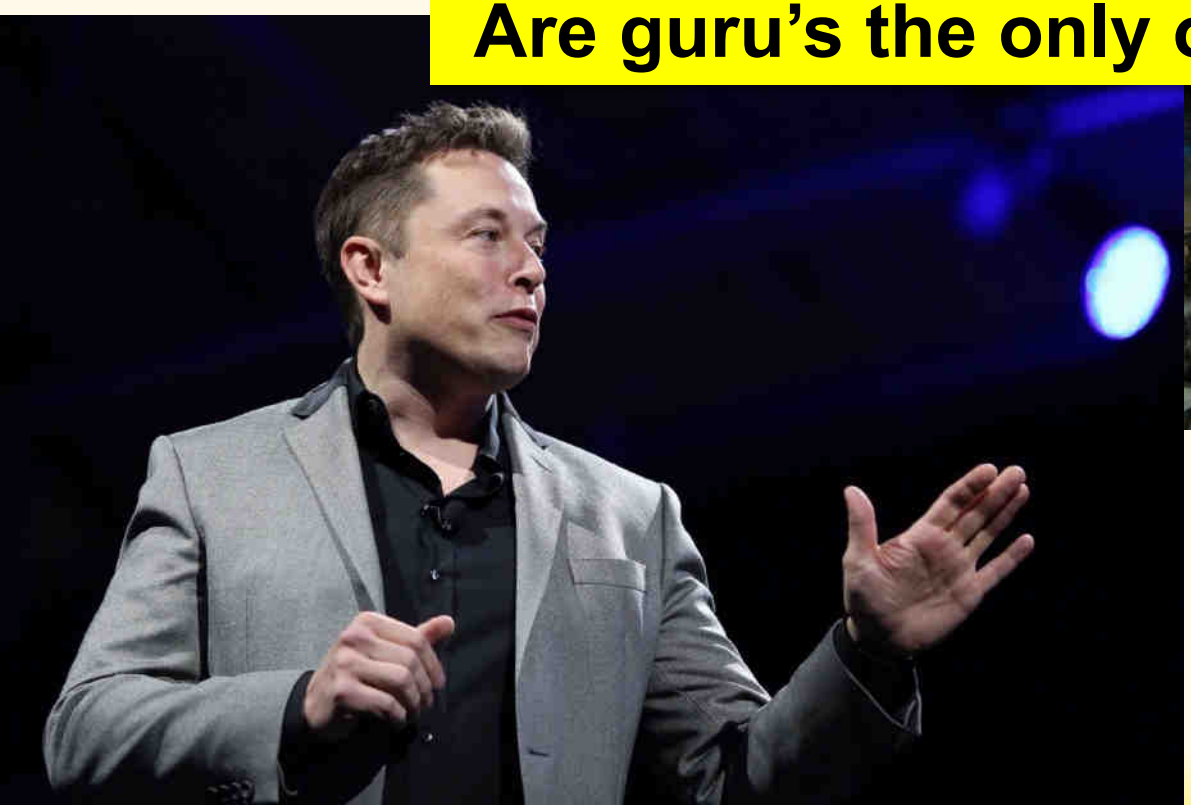
Radical / disruptive innovations

How to design (for)
radical / disruptive technologies?

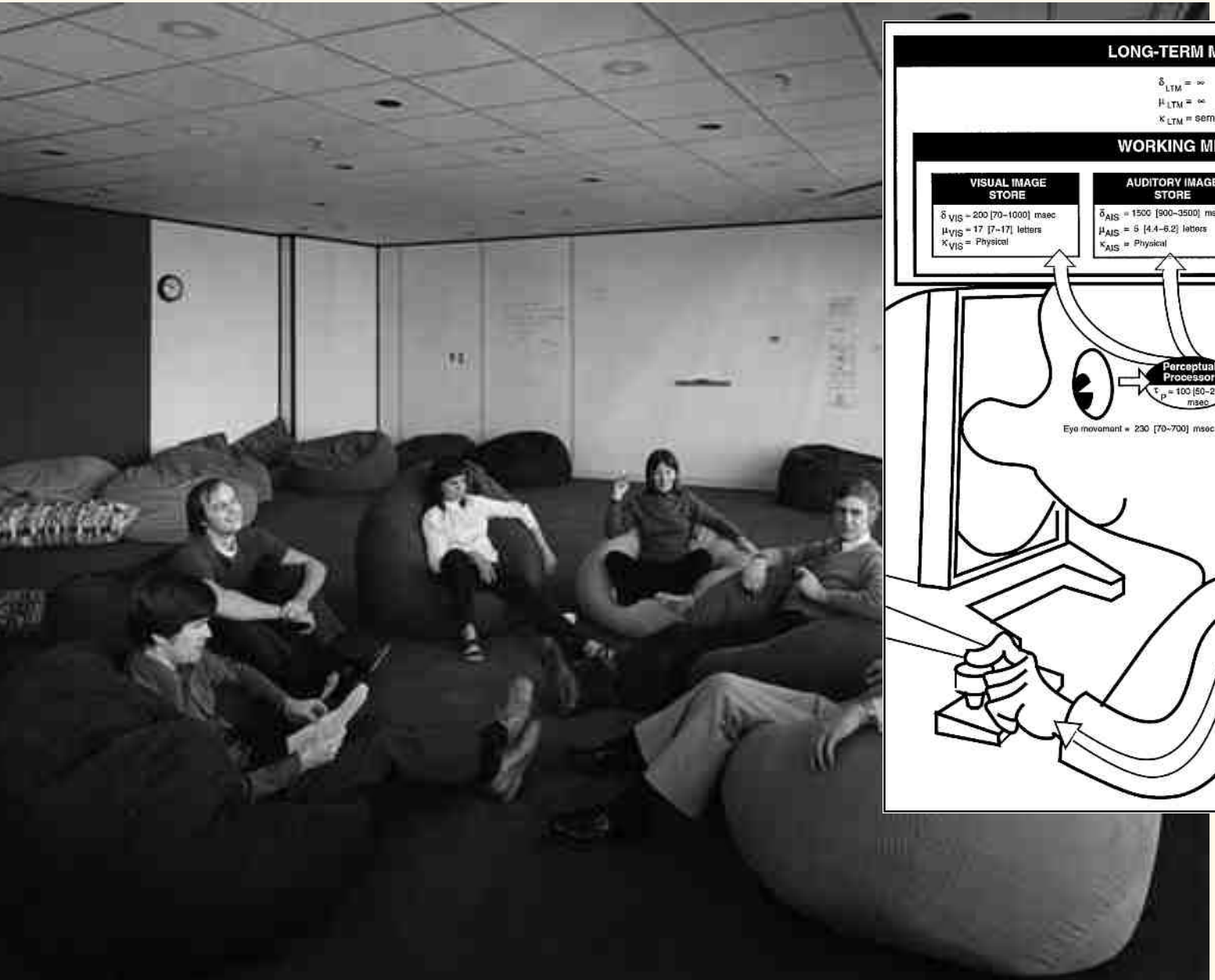
The guru approach



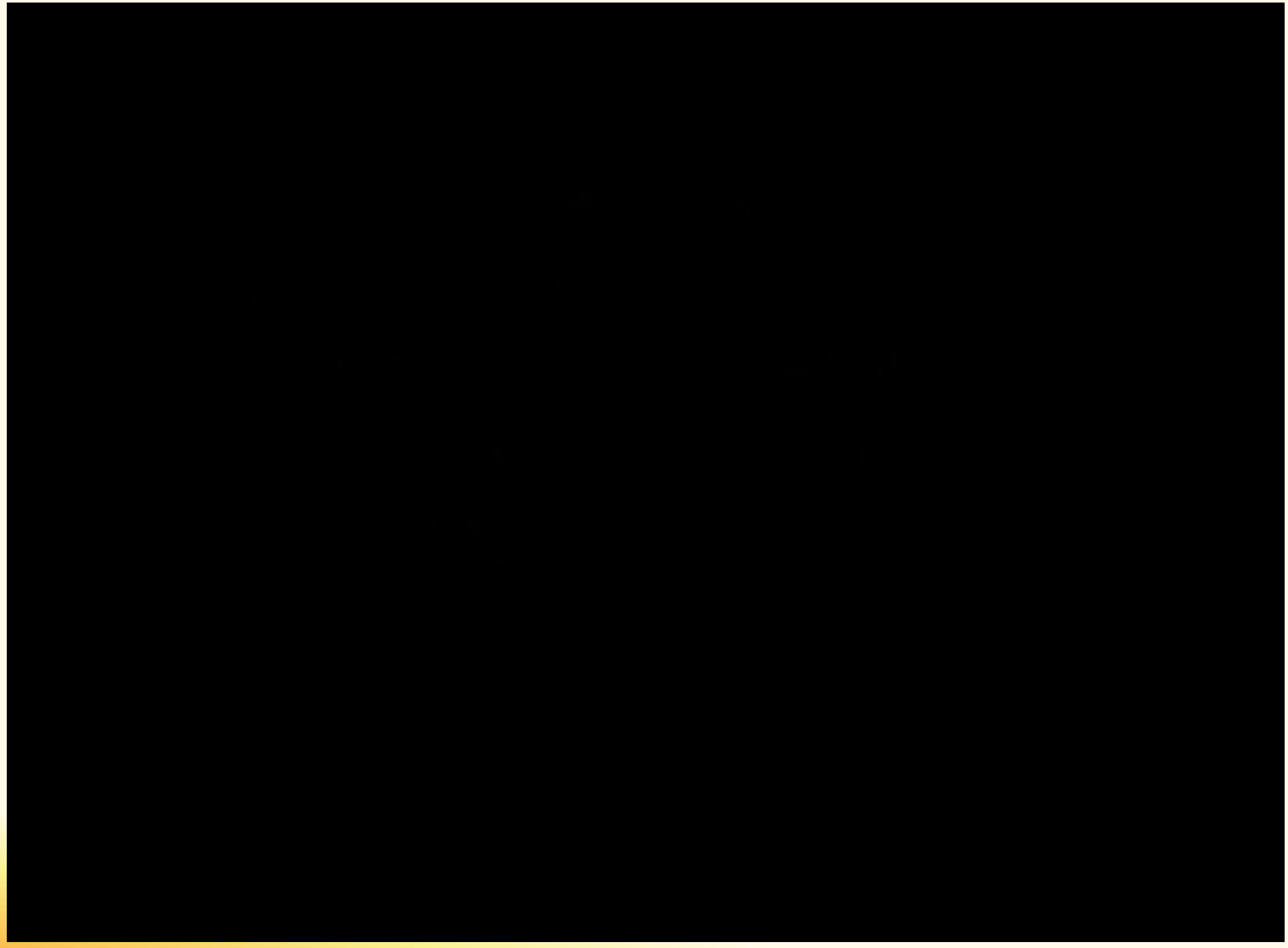
Are guru's the only option?



Science driven approach



Science driven approach



https://www.youtube.com/watch?v=DUwXN01ARYg

Science driven approach (incl science in industry)

- **Also thinking ahead, but..**
  - Grounded in principles and theories
  - Can make predictions beyond those in 1 guru's head
  - (most of the time) more rigorous
  - Out in the open ("open innovation")
- **Fruitful if:**
  - Able to think ahead of implications
  - Informed from multiple angles (inter- and multi-disciplinary)
  - Adaptive to changes
  - Through peer-review, debate, scrutiny

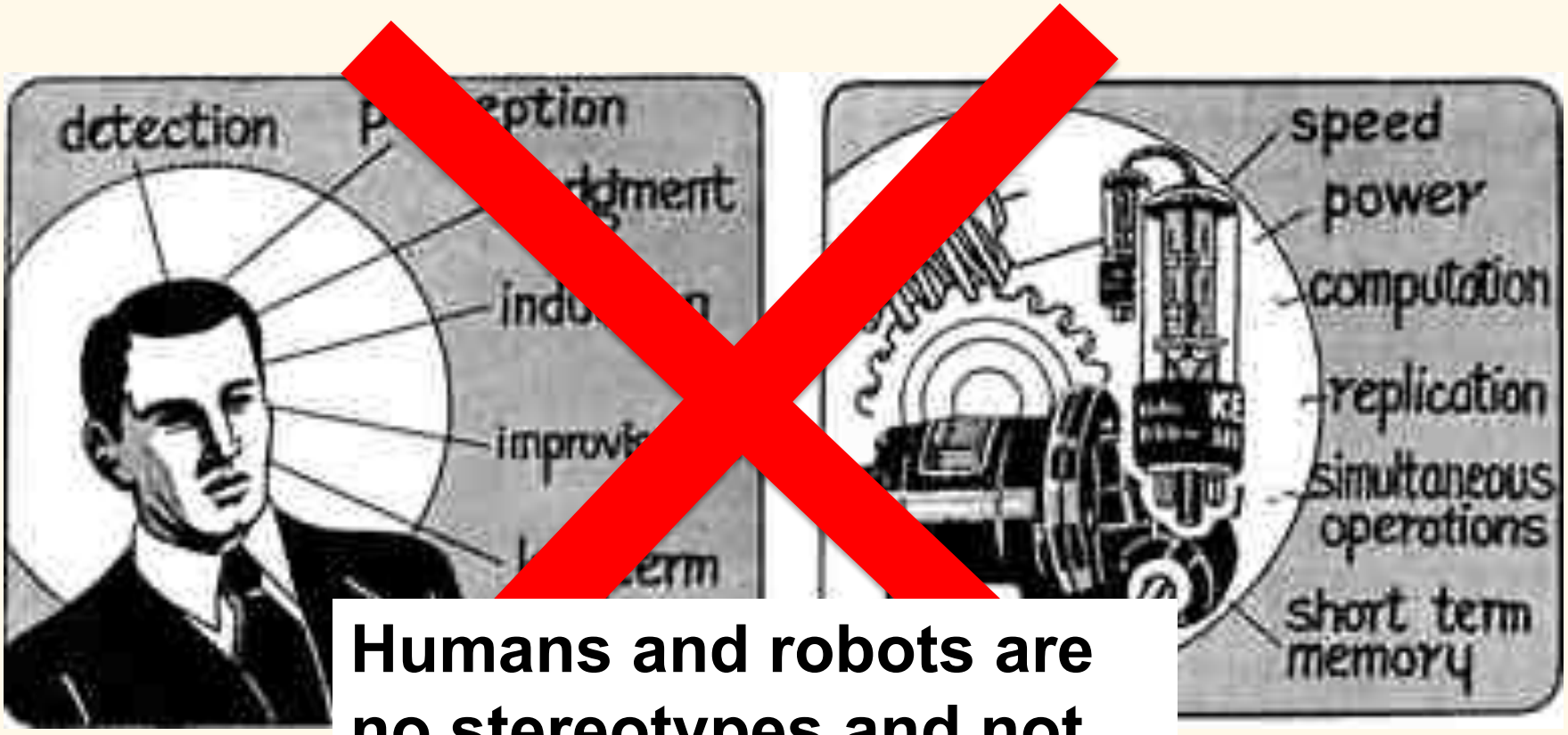
Today's topics

- Why consider design at all?
  - AI: Understanding by building
- Brief history of human-automation (and AI) interaction
- Radical / disruptive innovations
- **Why is designing for human-automation (and AI) interaction tricky?**
  - **Dynamic humans, systems, contexts, interaction**
- **Guidelines for design**
  - **Microsoft's guidelines for Human-AI Interaction**
  - **Lee, Wickens, Liu & Boyle's human-centric guidelines**

- **If you want to know more..**

Dynamic humans

Fitts (1951) list: MaBa-MaBa



Humans and robots are no stereotypes and not static in characteristics

Fitts PM (ed) (1951) Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC

Function (re-)allocation

Idea:

1. **Reduce tasks that humans can't do well**
2. **System will take over**
3. **Human can concentrate on remaining task and do better**

Reality:

3. Humans start doing other things

Well known:

Bainbridge (1983): Ironies of automation

Idea behind function (re-) allocation

Nissan concept: https://www.autofutures.tv/2019/05/22/nissan-unveil-worlds-first-next-gen-driver-assistance-system/

Reality

Accident with Uber, Tempe, AZ 18-03-2018

Not just 1 person

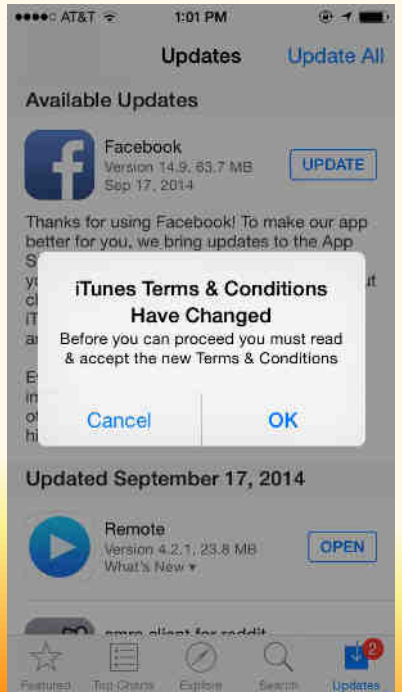
- **Meta-review by De Winter et al (2014) Transportation research F**
  - **Humans get more distracted when using vehicles with more automated features**

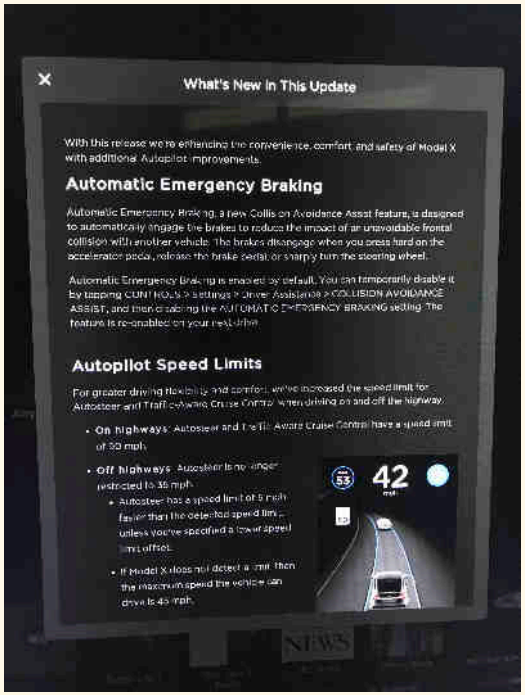
Dynamic humans

1. Irony of automation
   - Human behavior changes
   - Think also re-appropriation

Dynamic humans

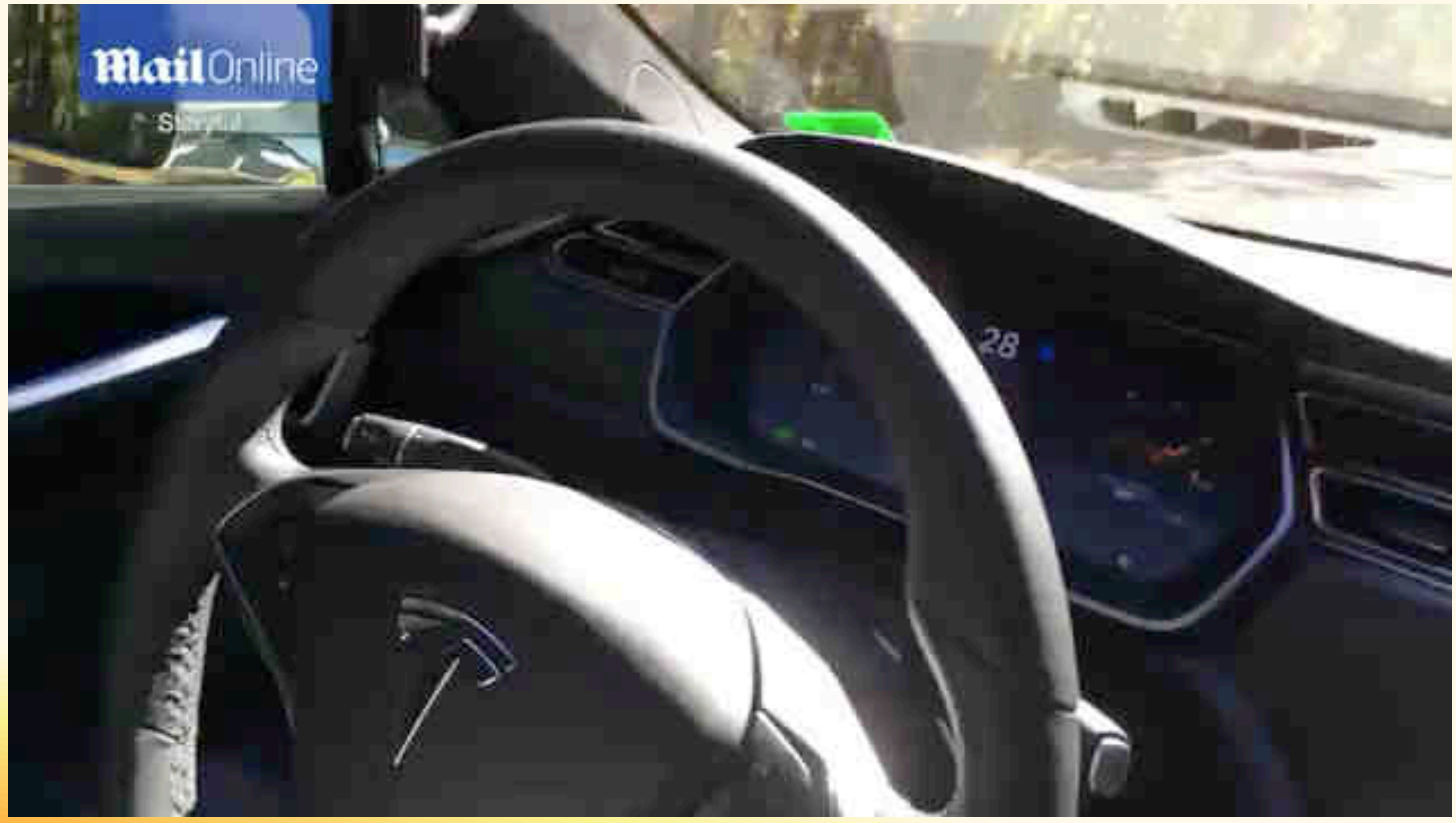
1. Irony of automation
   - Human behavior changes
   - Think also re-appropriation
2. (un-) learning of skill
   - Do we read the manual?



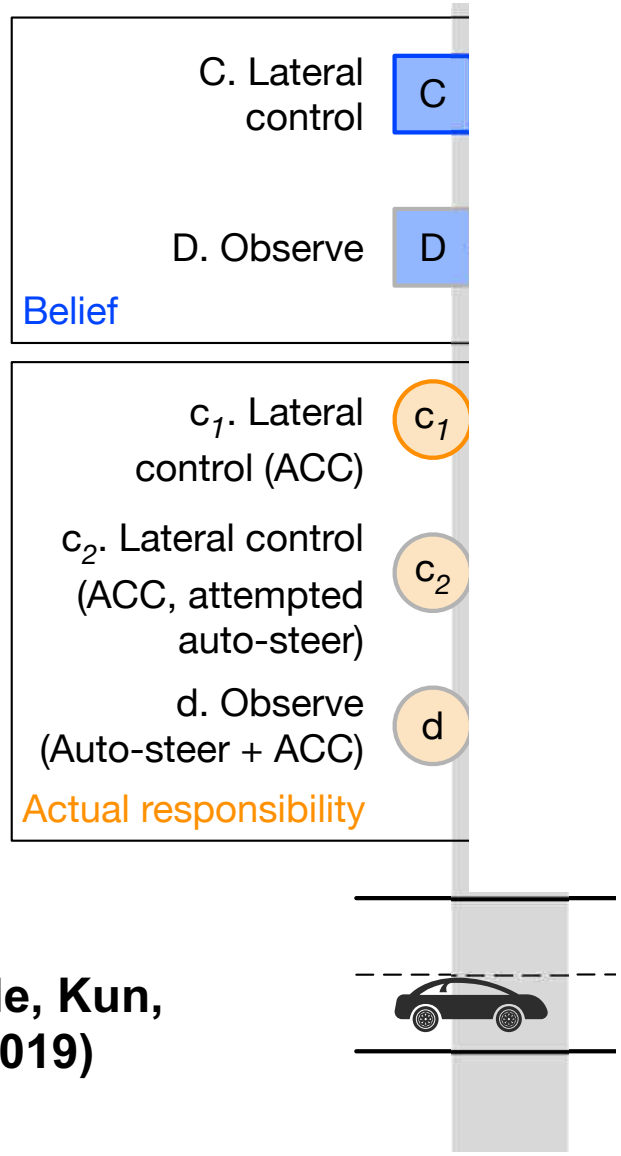


Dynamic machines

1. AI: Learning systems, system updates
2. Context of use / Operational design domain

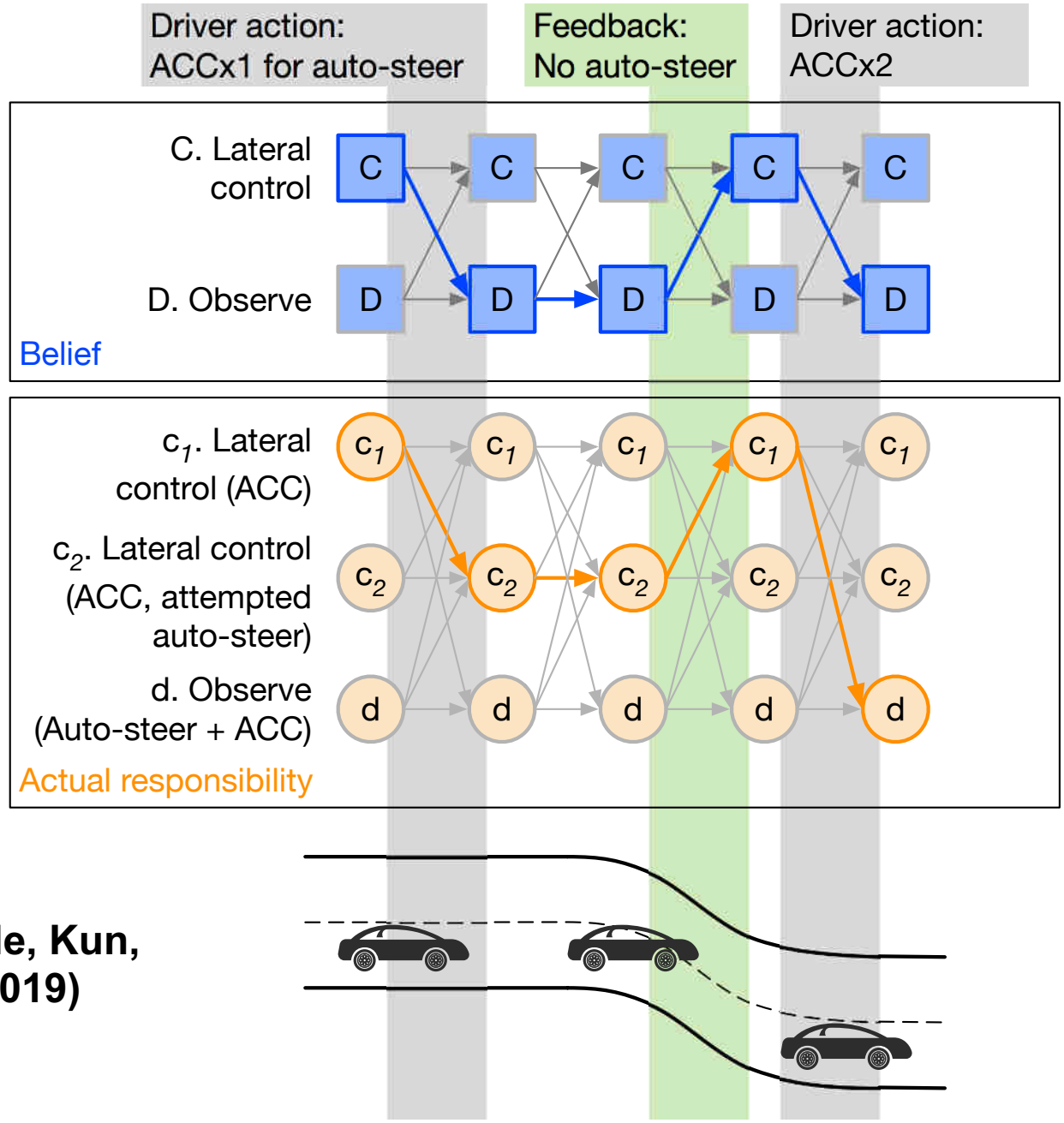


Dynamic contexts



Janssen, Boyle, Kun, Ju, Chuang (2019)

Dynamic humans, machines, contexts: mode confusion



Janssen, Boyle, Kun, Ju, Chuang (2019)

Dynamic interaction

- Due to dynamic:
  - Humans
  - Machines
  - Contexts
- Why hard to design?
  - Hard to anticipate all situations and scenarios

Today's topics

- Why consider design at all?
  - AI: Understanding by building
- Brief history of human-automation (and AI) interaction
- Radical / disruptive innovations
- Why is designing for human-automation (and AI) interaction tricky?
  - Dynamic humans, systems, contexts, interaction
- **Guidelines for design**
  - **Microsoft's guidelines for Human-AI Interaction**
  - **Lee, Wickens, Liu & Boyle's human-centric guidelines**

- **If you want to know more..**

Guidelines for design

The Guidelines for Human-AI Interaction will help you create AI systems and features that are human-centered. We hope you use them throughout your design process – as you evaluate existing ideas, brainstorm new ones, and collaborate with the multiple perspectives involved in creating AI.

These guidelines synthesize more than 20 years of thinking and research in human-AI interaction. Learn more: https://aka.ms/aiguidelines

Guidelines for design (Microsoft paper)

- **Read the paper yourself**
  - **Interesting guidelines**
  - **Interesting method for how to create list**
  - **Interesting way of testing**
  - **Interesting examples**

See online website for more material, that might be useful for your future job: https://aka.ms/aiguidelines

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S.T., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM. Paper 3.

Core message

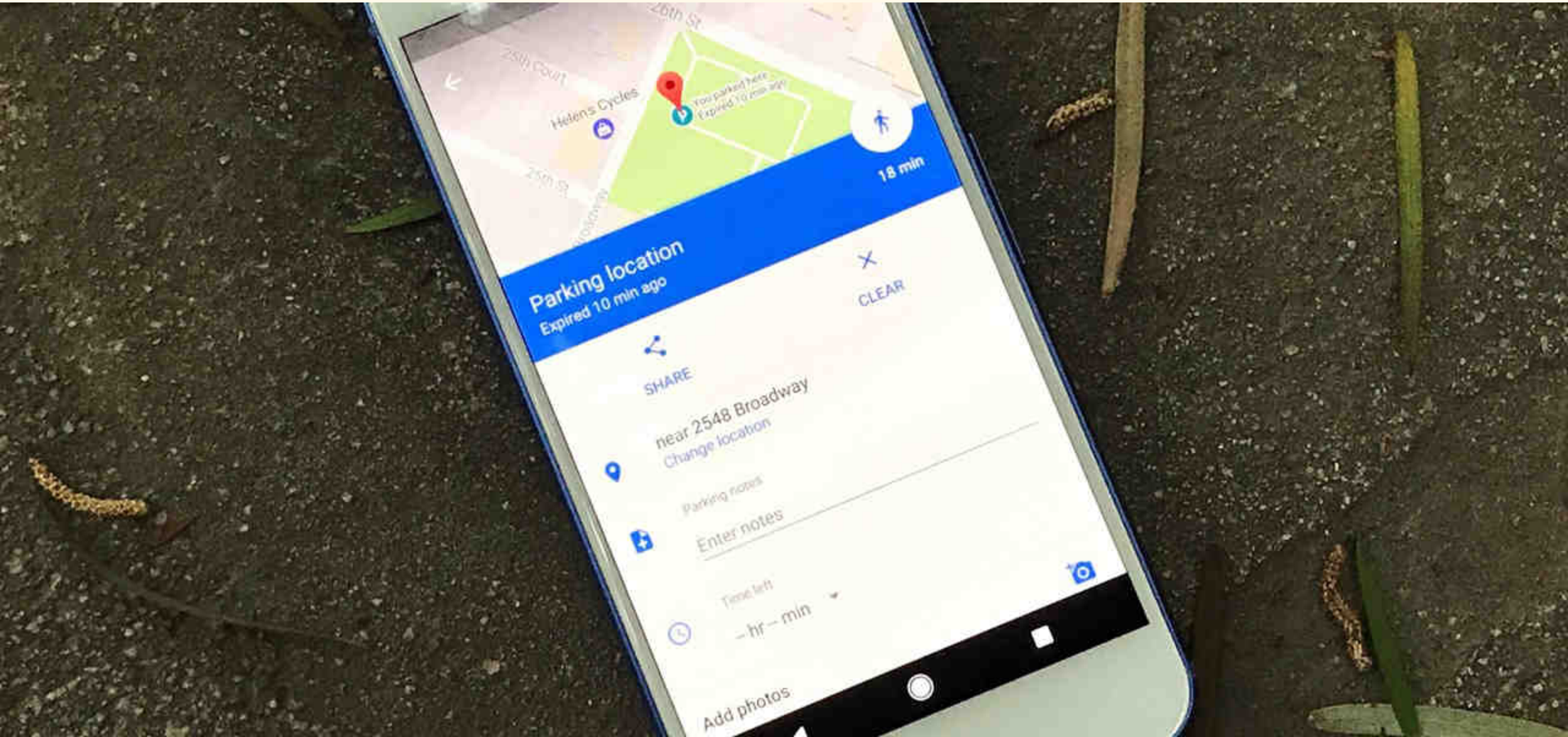
- **Think about design in all phases of product R&D + employment**

Core message

- **Use for design:**
  - **How can we accommodate this guideline?**
  - **How can we overcome associated challenges?**
- **Use for evaluation:**
  - **Has the system met the guideline, and how?**
  - **Where has the system not met the challenge, and how?**

Example: guideline 4

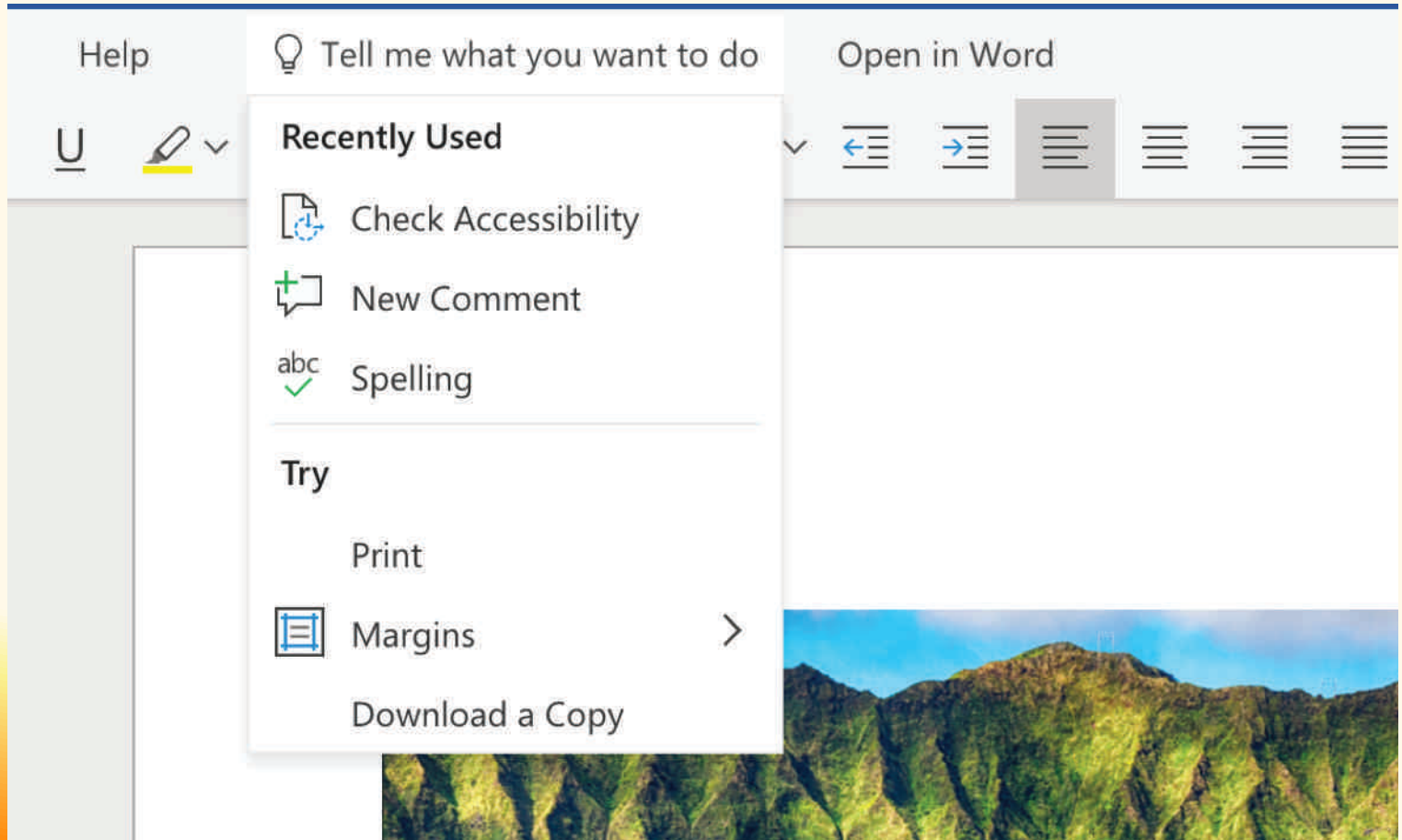
Show contextually relevant information.
Display information relevant to user's current task and environment



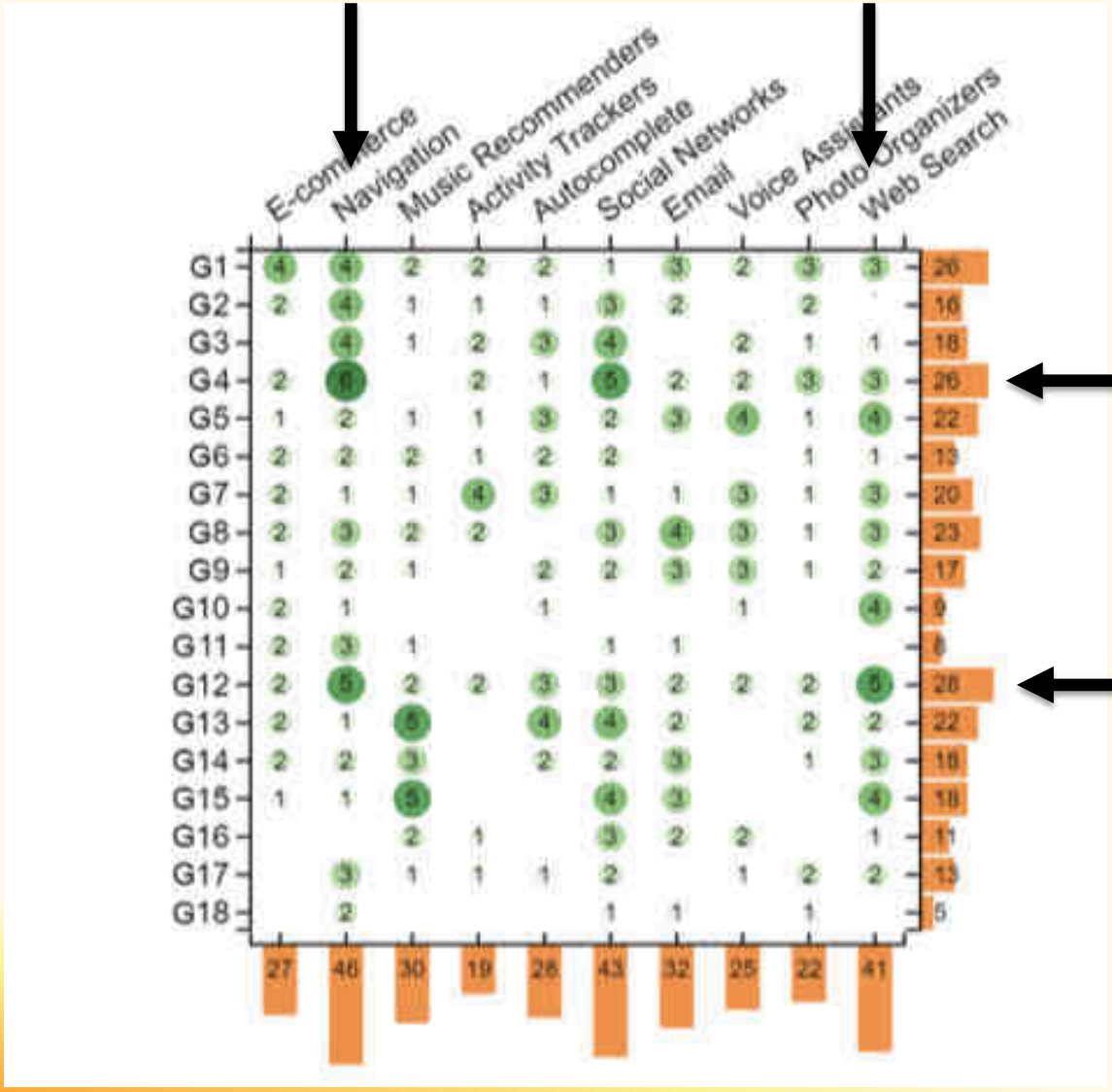
Example: guideline 13

Learn from user behavior

Personalize the user's experience by learning from their actions over time

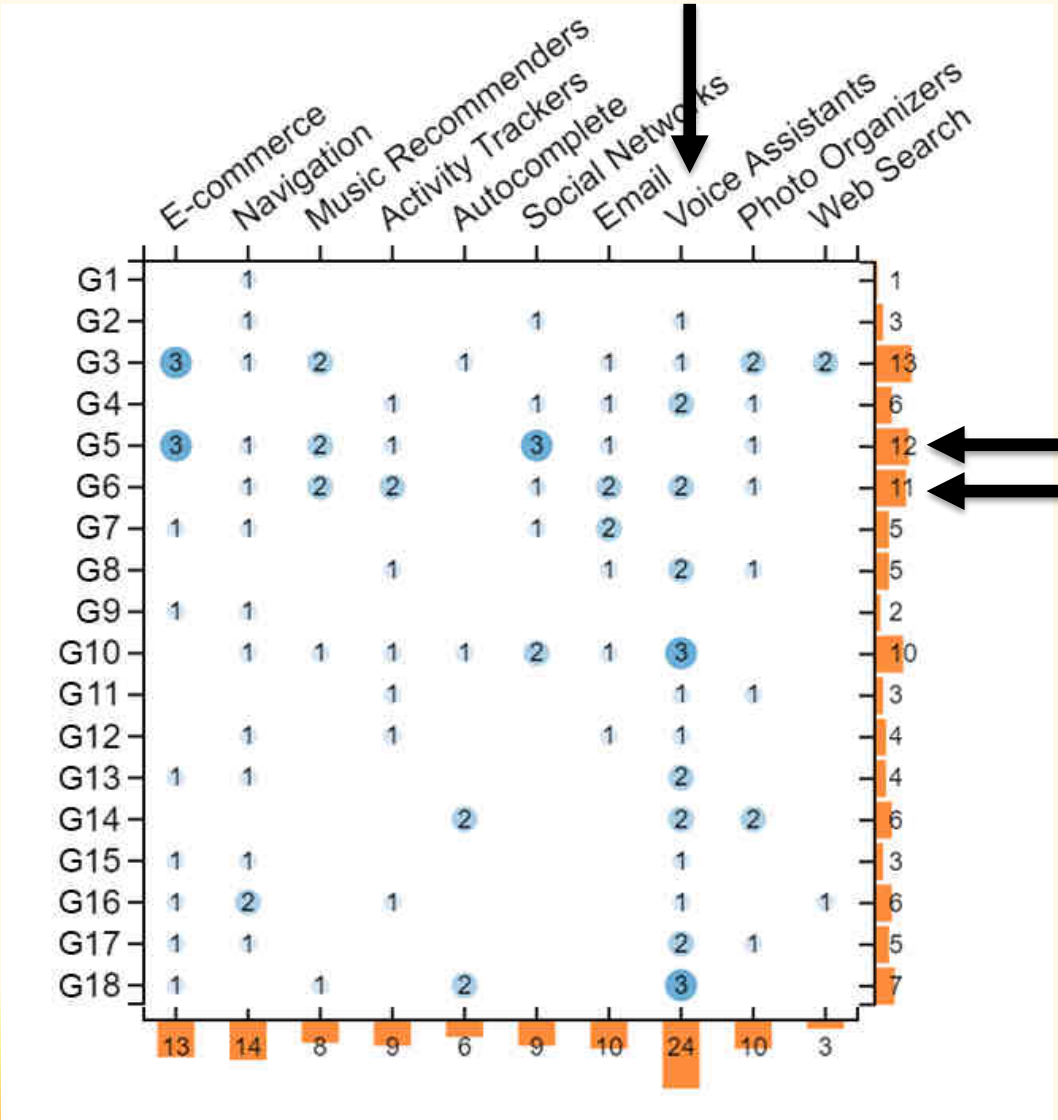


What works in practice?



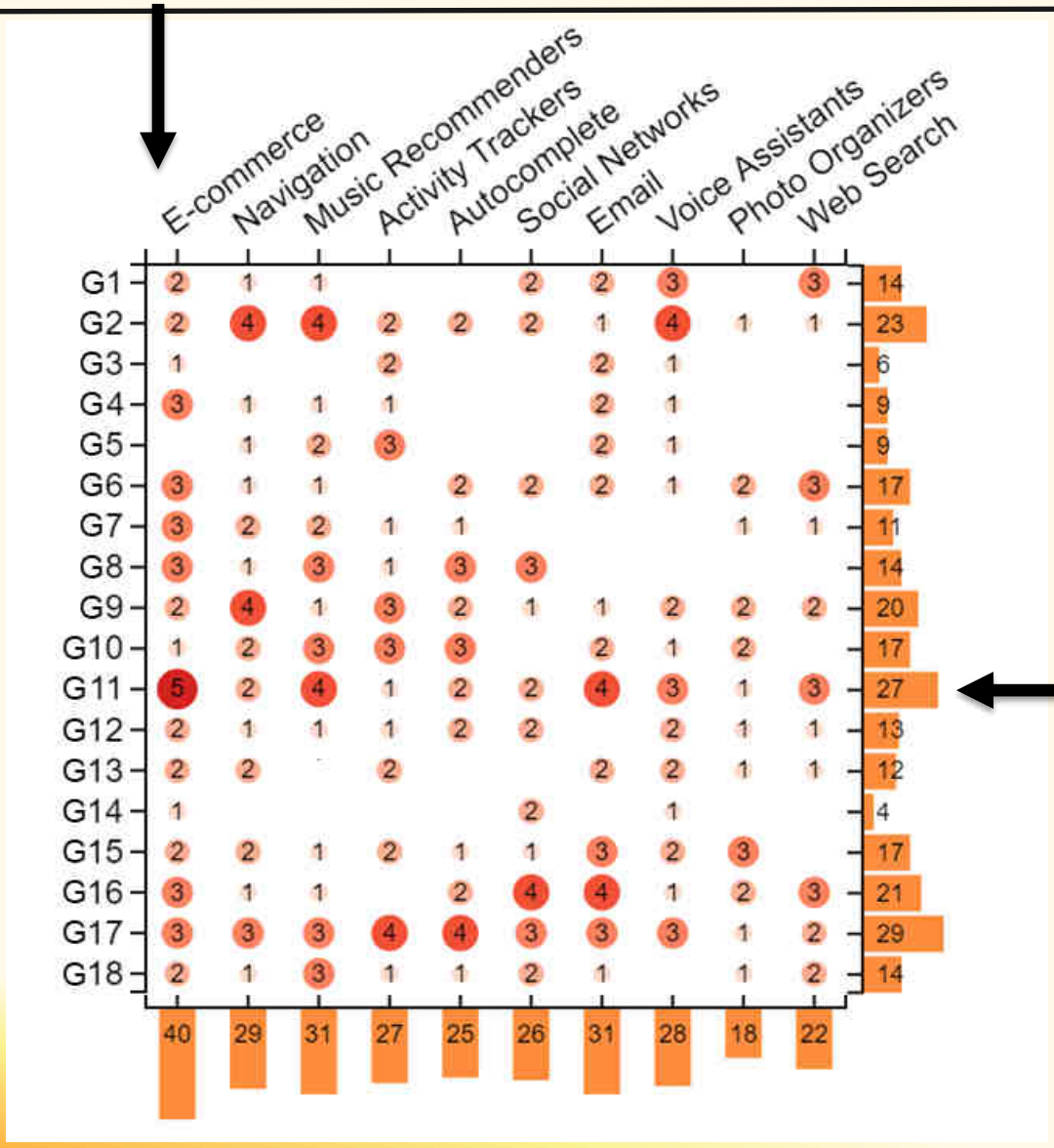
(according to experts)

Which rules do not apply



(according to experts)

Which rules are violated

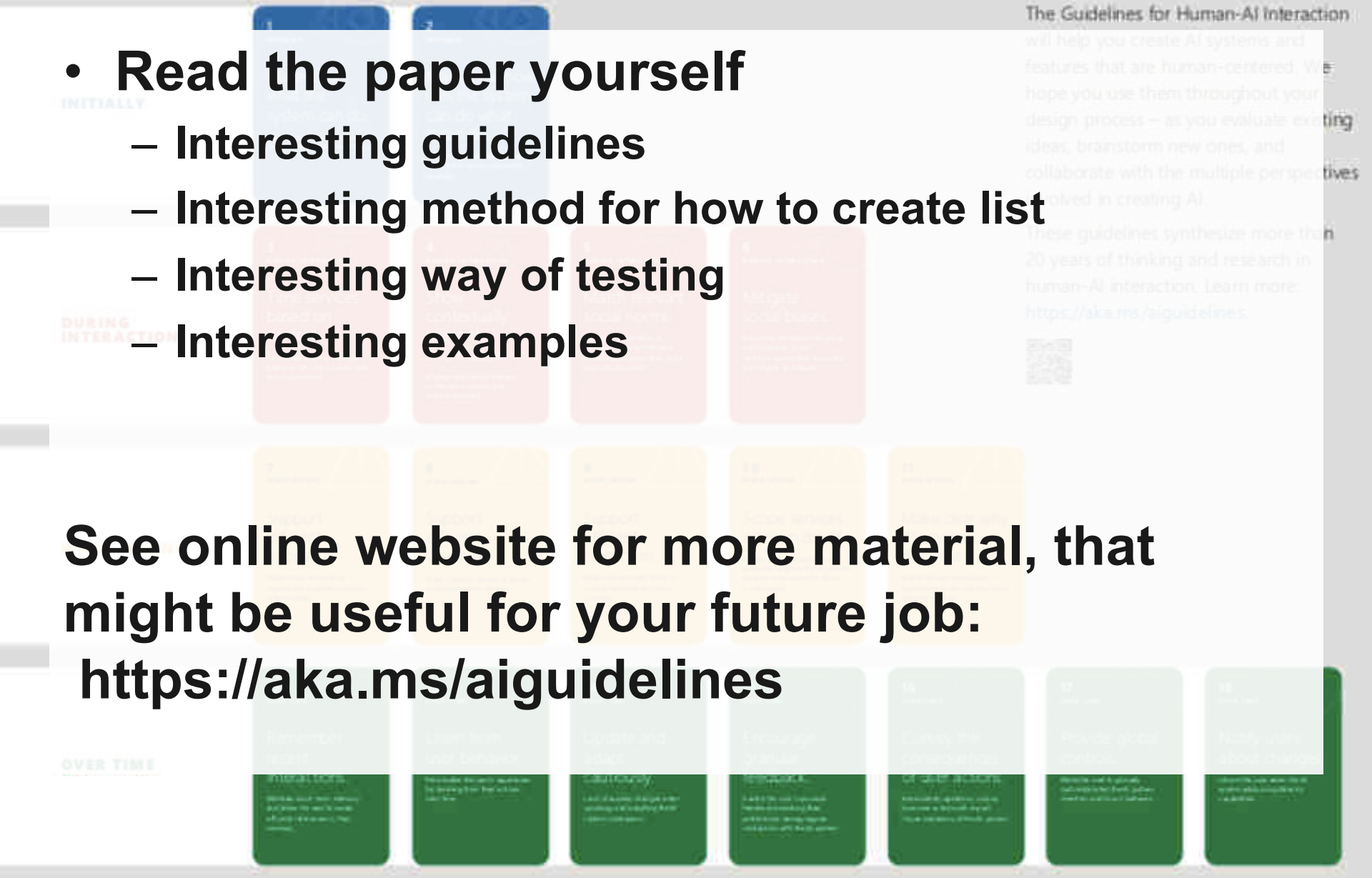


(according to experts)

Guidelines for design

- **Read the paper yourself**
  - **Interesting guidelines**
  - **Interesting method for how to create list**
  - **Interesting way of testing**
  - **Interesting examples**

See online website for more material, that might be useful for your future job:
https://aka.ms/aiguidelines



Framework Lee, Wickens, Liu, Boyle

Lee, J. D., Wickens, C. D., Liu, Y., & Boyle, L. N. (2017). Human-Automation Interaction. In *Designing for People: An Introduction to Human Factors Engineering*. Charleston, SC: CreateSpace.

Focus Lee et al:
"what *aspects/characteristics* of *people* should designer/research consider"

Compared to microsoft paper:
"what is an *action* that a *designer* needs to take"

Designing for people

- **Mental model principles**
- **Attention principles**
- **Perception principles**
- **Response selection principles**
- **Interaction principles**
- **Organizaional principles**

(Model of) human (understanding) driven

➔ Relevant for Utrecht AI researchers (human-centered)

Example: Perception principles

- **P7: Transparency – keep the person informed**

All principles from Lee et al.

Mental model principles
MM1. Define and communicate the purpose of automation
MM2. Define and communicate the operating domain
MM3. Design the role of the person and automation
MM4. Simplify the mode structure MM5. Make trustable and polite

Attention principles
A6. Signal inability to satisfy role

Perception principles
P7. Transparency–keep the person informed

Response selection principles
R8 Avoid accidental activation and deactivation

Interaction principles
I9. Keep the person in the loop
I10. Support smooth re-entry into the loop
I11. Make automation directable
I12. Make automation flexible and adaptable
I13. Consider adaptive automation

Organizational principles
O14. Keep people trained
O15. Consider organizational consequences

Today's topics

- **Why consider design at all?**
  - **AI: Understanding by building**
- **Brief history of human-automation (and AI) interaction**
- **Radical / disruptive innovations**
- **Why is designing for human-automation (and AI) interaction tricky?**
  - **Dynamic humans, systems, contexts, interaction**
- **Guidelines for design**
  - **Microsoft's guidelines for Human-AI Interaction**
  - **Lee, Wickens, Liu & Boyle's human-centric guidelines**

- **If you want to know more..**

If you want to know more

- Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. *International Journal of Human-Computer Studies, 31,* pp99-107. https://www.sciencedirect.com/science/article/pii/S1071581919300552
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S.T., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM. Paper 3. Paper and extra material at: https://aka.ms/aiguidelines
- Lee, J. D., Wickens, C. D., Liu, Y., & Boyle, L. N. (2017). Human-Automation Interaction. In *Designing for People: An Introduction to Human Factors Engineering*. Charleston, SC: CreateSpace.

Questions?

Chris Janssen

c.p.janssen@uu.nl
www.cpjanssen.nl