Practical Methods in Human-Centred AI

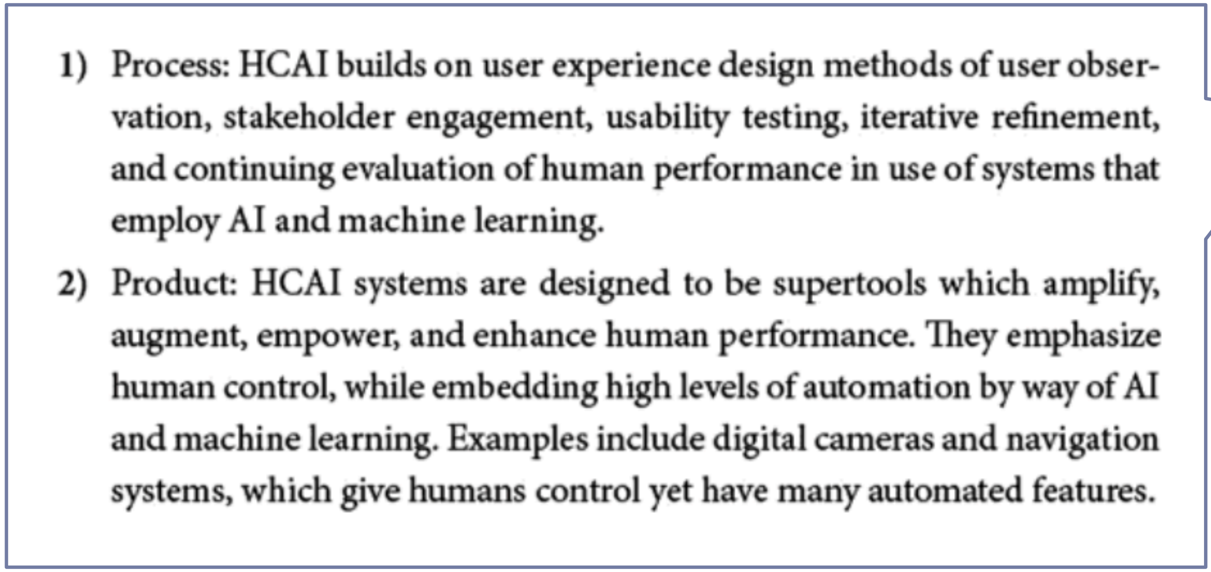
ACS P342 / Part II unit - Alan Blackwell & Advait Sarkar

Overview

- Practical experimental course
  - lectures provide overview and sample of current research
- This introduction
  - general principles, research approaches, strategic trends
- Specialist lectures
  - six deep-dive topics, each illustrating some practical methods
- Design and run your own study
  - practical feedback on work in progress every week
- Final presentation of your results

Course objective

- "Human-Centered AI"
  Ben Shneiderman
  (OUP 2022)



1) Process: HCAI builds on user experience design methods of user observation, stakeholder engagement, usability testing, iterative refinement, and continuing evaluation of human performance in use of systems that employ AI and machine learning.
2) Product: HCAI systems are designed to be supertools which amplify, augment, empower, and enhance human performance. They emphasize human control, while embedding high levels of automation by way of AI and machine learning. Examples include digital cameras and navigation systems, which give humans control yet have many automated features.

The book of the course (in Alan's mind – Advait has views!)

Moral Codes

Designing software without surrender to AI

Alan Blackwell
(MIT Press 2024)

Free preview:
https://moralcodes.pubpub.org

Where Ben, Alan and Advait agree with Google DeepMind

- Four waves of AI, according to DeepMind founder Demis Hassabis:
  - First wave (GOFAI): Expert systems & symbolic reasoning
  - Second wave: Statistical inference
  - Third wave: Deep learning
  - Fourth wave: Intelligent tools
- Our approach:
  - Intelligent tools as advanced HCI
  - Including: Visualisation, Programming, Labelling, Explanation, Predictive Text
- A *practical* HCI course:
  - Project work to build, customise, measure, observe …
- For: Part III and MPhil ACS (research preparation), Part II (advanced HCI)

Your background

- 1. Prior HCI experience
- 2. Prior ML/AI experience
- 3. What do you hope to get out of this course?

	None	Casual	Student	Professional
HCI	4	2	9	5
ML	1	2	8	8

Target outcome

- This is a specialised and focused practical research training course.
- The expected outcome:
    - You will achieve research competence in a recognised academic field such as Intelligent User Interfaces, Interactive Intelligent Systems etc
- ACS assessment will be relative to the international standard of graduate students working in these fields.
    - Written work will be graded relative to typical student publications in the field
    - Presentations will be expected to meet the standard of first-year PhD students in the field, for example at the Doctoral Consortium of a specialised conference.
- Part II students demonstrate skills by "replicating" a competent study.

Lecture topics

- Week 5 – Labelling as a fundamental problem (AS)
  - attribution, subjectivity, reliability, consistency
- Week 3 - Mixed initiative interaction (AB)
  - information gain, cognitive ergonomics, agency & control
- Week 4 - Program synthesis (AB)
  - end-user programming, attention investment
- Week 2 - Generative AI (AS)
  - (under development!)
- Week 6 – Bias and fairness (AB)
  - discrimination, accountability and ethics in hybrid systems
- Week 7 - Explainability (guest)
- Week 8 – Your research presentations

Practical work plan

- Week 1 - select research question
- Week 2 - discuss potential study approaches
- Week 3 - review and feedback on study proposals
- Week 4 & 5 - review logistical issues / practical progress
- Week 6 - discuss preliminary findings
- Week 7 - discuss research implications
- Week 8 - final presentation

Assessment for ACS

- Final research report (80%)
  - Based on your practical work
  - Presented as an original research paper
- Optional (but recommended) work-in-progress drafts
  - Advisory grades will be provided as feedback, for revision in final report
- Reflective diary (20%)
    - Summarise lectures
    - Document discussions
    - Record development of your own thinking
    - Make 8 weekly entries …
    - … bind together and submit with a final summative review

Assessment for Part II

- Final research report (80%)
  - Based on your practical work
  - Presented as a research paper replicating a previous publication
- Ticks awarded for work-in-progress drafts (20%)
  - Advisory grades will be provided as feedback, for revision in final report

Practical work-in-progress

- Week 2 - Research question (200 words) + a sample diary entry for ACS
- Week 3 - Study design (400 words)
- Week 4 - Another sample diary entry for ACS
- Week 5 - Draft literature review for final report (400 words)
- Week 6 - Draft introduction to report (200 words)
- Week 7 - Draft results section for report (400 words)
- Week 8 - Draft discussion section for report (200 words)

“Indicative feedback” on work in progress

- A+ excellent - on target for 85-100
- A very good - on target for 75-85
- B good - on target for 70-80
- C acceptable - on target for 60-70
- D disappointing - risk of fail

- The final grade will be awarded solely on the basis of the final report, and you are welcome to change as much as you like in response to feedback, or to simply copy draft material straight in, whichever you prefer.

Reading suggestions

- Refresh knowledge of undergraduate HCI
  - Cambridge lecture notes (and YouTube videos) for *Further HCI*
    - (refresher: Research Skills unit this Friday morning
  - Preece, Rogers and Sharp *Interaction Design beyond HCI*
- Blackwell (2024)
  - *Moral Codes*
- Review Cambridge guidance on human participants
  - https://www.tech.cam.ac.uk/research-ethics/school-technology-research-ethics-guidance
- Cairns and Cox (2008)
  - *Research Methods for Human-Computer Interaction*
- Carroll (2003)
  - *HCI Models, Theories and Frameworks*
- **Mostly: Recent research literature**

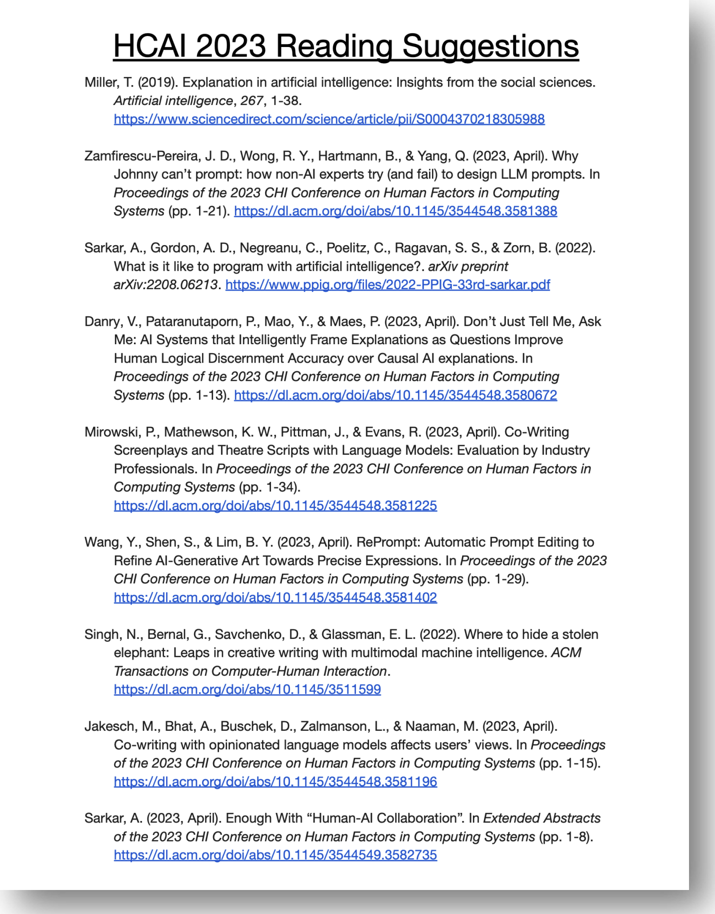
A note about the reading list

Available on course materials page.

Don’t try to read all of it!

Chosen because:
- Influential
- Well-executed research
- Interesting/unique angle

Papers on this list may be suitable as a basis for your own research question/study design.



HCAI 2023 Reading Suggestions

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38. https://www.sciencedirect.com/science/article/pii/S0004370218305988

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can’t prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-21). https://dl.acm.org/doi/abs/10.1145/3544548.3581388

Sarkar, A., Gordon, A. D., Negreanu, C., Poelitz, C., Ragavan, S. S., & Zorn, B. (2022). What is it like to program with artificial intelligence?. *arXiv preprint arXiv:2208.06213*. https://www.ppig.org/files/2022-PPIG-33rd-sarkar.pdf

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023, April). Don’t Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). https://dl.acm.org/doi/abs/10.1145/3544548.3580672

Mirowski, P., Mathewson, K. W., Pittman, J., & Evans, R. (2023, April). Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-34). https://dl.acm.org/doi/abs/10.1145/3544548.3581225

Wang, Y., Shen, S., & Lim, B. Y. (2023, April). RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-29). https://dl.acm.org/doi/abs/10.1145/3544548.3581402

Singh, N., Bernal, G., Savchenko, D., & Glassman, E. L. (2022). Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*. https://dl.acm.org/doi/abs/10.1145/3511599

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023, April). Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). https://dl.acm.org/doi/abs/10.1145/3544548.3581196

Sarkar, A. (2023, April). Enough With “Human-AI Collaboration”. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-8). https://dl.acm.org/doi/abs/10.1145/3544549.3582735

Theories of interaction

Human-Computer Interaction (HCI) - Three waves

- First wave (1980s):
  - Theory from Human Factors, Ergonomics and Cognitive Science
- Second wave (1990s):
  - Theory from Anthropology, Sociology and Work Psychology
- Third wave (2000s):
  - Theory from Art, Philosophy and Design

First wave: HCI as engineering "human factors" (1980s)

- The "user interface" (or MMI "man-machine interface") was considered to be a separate module, designed independently of the main system.
- Design goal was efficiency (speed and accuracy) for a human operator to achieve well-defined functions.
- Use methods from cognitive science to model the user's perception, decision and action processes, and predict usability on the basis of that model
  - At this point, relatively closely aligned with AI

Second wave: HCI as social system (1990s)

- AI models did not result in more usable machines (see esp. Lucy Suchman)
  - Resulted in a significant intellectual challenge to cognitive science and AI!
- The design of complex systems is a socio-technical experiment
  - Took account of other information factors including conversations, paper, and physical settings
- Study the context where people work
  - Used ethnography (or "Contextual Inquiry" or "Workplace Studies") to understand other ways of seeing the world and characterise social structures
- Other stakeholders are integrated into the design process
  - Prototyping and participatory workshops aim to empower users and acknowledge other value systems

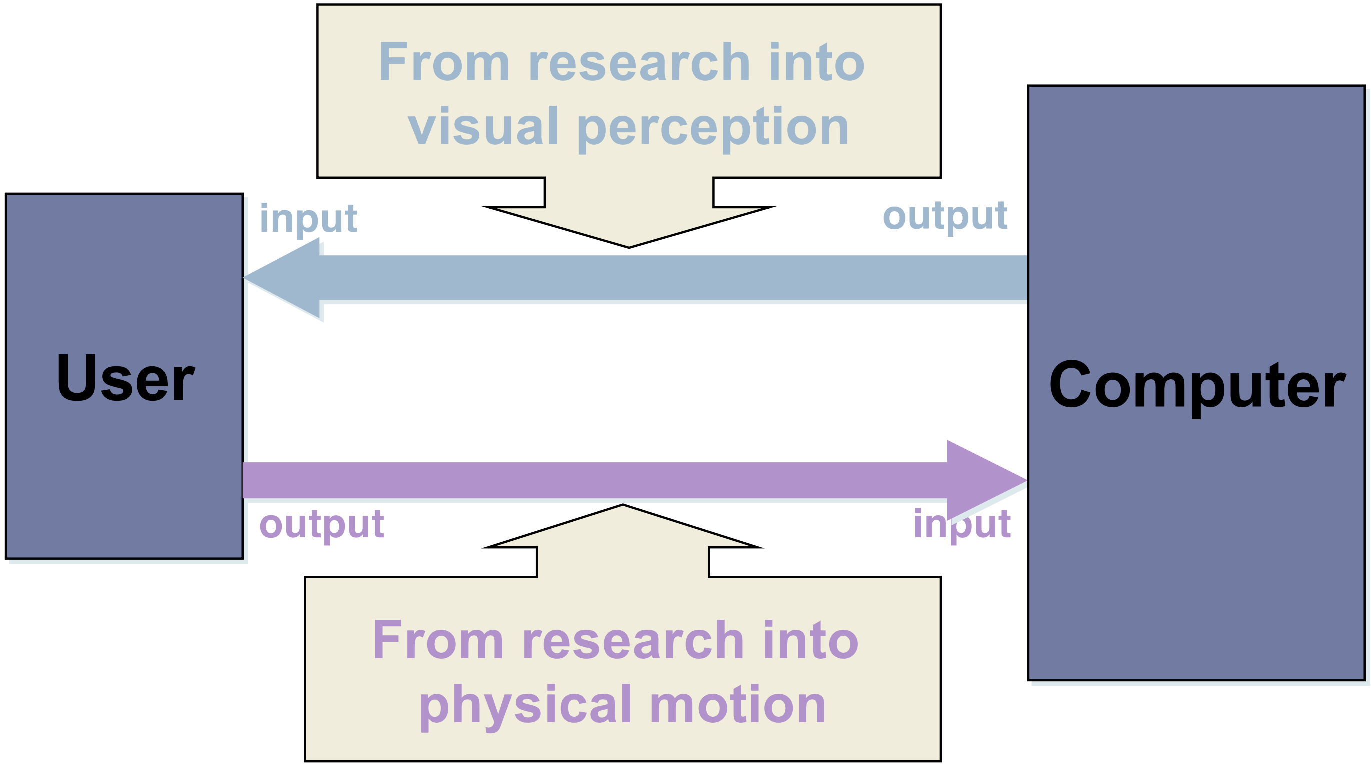
Third wave: HCI as culture and experience (2000s)

- Ubiquitous computing affects every part of our lives
  - It mixes public (offices, lectures) and private (bedrooms, bathrooms)
- Outside the workplace, efficiency is not a priority
  - Usage is discretionary
  - User Experience (UX), includes aesthetics, affect,
- Design experiments are speculative and interpretive
  - Critical assessment of how this is meaningful
- Was until 2018 pretty much completely divorced from AI
  - But this is changing very rapidly, as critical AI studies mature!

Summary of Cambridge HCI content

- Textbooks
  - Preece, Sharp & Rogers
  - Carroll
- Part 1a Interaction Design
  - Requirements analysis and design process, data collection (observation, interviews, focus groups) and analysis. Design and prototyping, personas, storyboards and task models. Principles of good design. Human cognition. Usability evaluation.
- Part 1b Further HCI
  - Theory driven approaches. Design of visual displays. Goal-oriented interaction. Designing smart systems. Designing efficient systems. Designing meaningful systems. Evaluating interactive system designs. Designing complex systems.
- Part 2/3
  - Affective Computing, Computer Music (not in 2023/24), Advanced Graphics …

Cognitive neuroscience and first-wave HCI



From research into visual perception

input

output

User

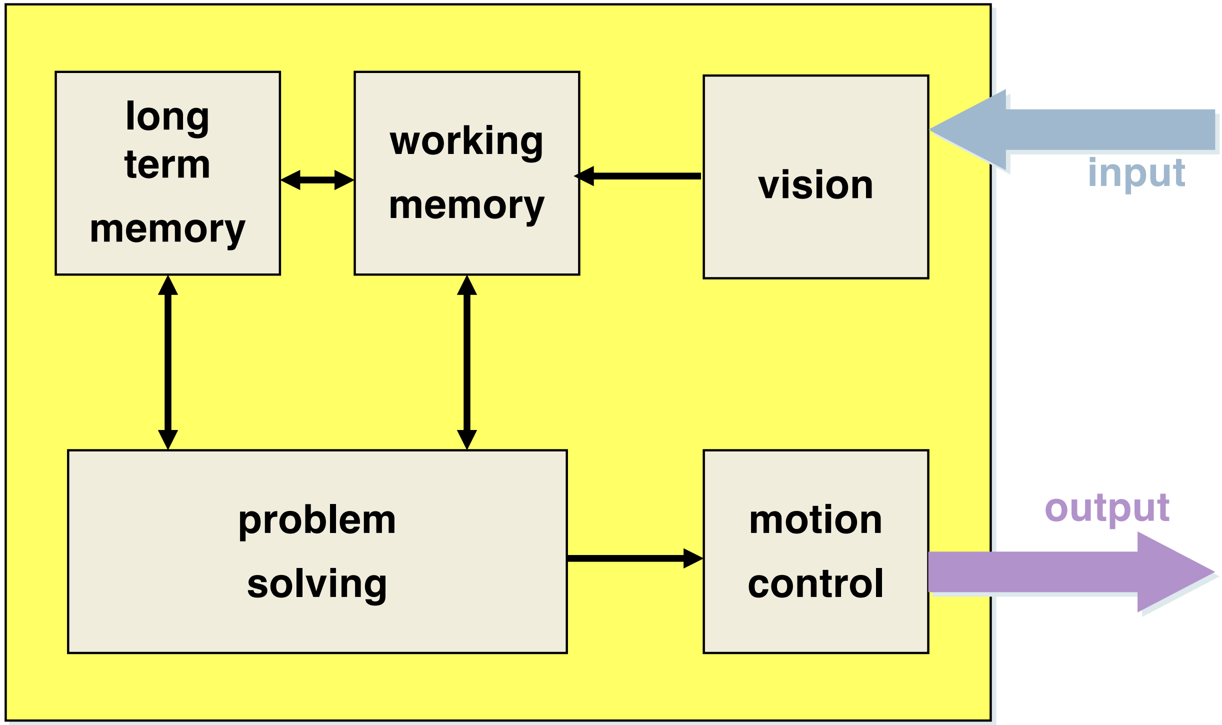
Computer

output

input

From research into physical motion

Neuroscience as computational user ‘boxology’



Engineering models of human I/O, memory, CPU

- Seeks "impedance match" of computer with computational user model
  - Extend principles of human factors and ergonomics
  - Psychophysical perception
  - Speed and accuracy of movement at keystroke level
  - Measure reaction time (and infer decision time?)
  - Include working memory capacity
    - 7 +/- 2 'chunks'
    - Single visual scene
  - GOFAI-planner style Goals Operators Methods Selection
- Is intelligent task design a matter of 'cognitive ergonomics'?

The problem of (human) learning

- Classical models assumed users would be *made* to read the manual
- In contrast, *discretionary usage* systems require exploratory learning models because users can (and do) walk away
  - Focus on minimal instruction, immediate progress toward user goals
  - Now taken for granted (but only after long battle with usability advocates)
- Cognitive walkthrough review methods allowed system designers to anticipate usability problems, based on model of situated learning rather than cognitive model of planning

The sticky problem of viscosity

- Deciding what to do is often harder than doing it
  - But HCI models assume a ‘correct’ sequence of actions
- How do you change your mind if something goes wrong?
  - problem solving
  - planning
  - knowledge representation
- External representations are often required
  - But did the designers anticipate people making mistakes?
- Many systems and visual representations make it hard to change your mind, or to engage in exploratory design
  - Complex systems can be regarded as interaction spaces

Wicked problems (Rittel & Webber)

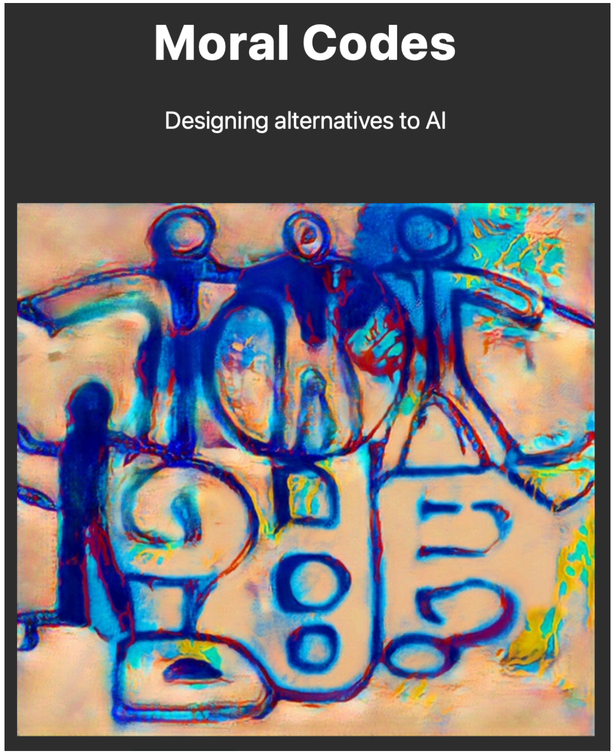
- Formulated in reaction to promotion of AI/cybernetic methods (e.g. optimization, goal-directed search) in business schools and public policy
- Wicked problems have:
  - no definitive formulation
  - no stopping rule
  - no true-or-false outcome: only good-or-bad
  - no ultimate test of a solution
  - no set of permissible operations
  - essentially unique

Older paradigms of intelligent user interfaces

- Perfect information games (toy worlds, chess, go, videogames)
  - Not considered particularly interesting
- Recommender systems
  - Once a major research area, now familiar - Amazon, Spotify, YouTube, Netflix, etc.
- Scripted dialogue / heuristic-based chatbots and agends
  - e.g. voice assistants – but watch "guardrails" become recommenders!
- Programming by example, program synthesis
  - See Lieberman *Watch What I Do*, but also e.g. Microsoft Excel FlashFill
  - Advances in code generation: codex, github copilot
- Human-in-the-loop automation
  - Autopilots, remote-operation, "autonomous" vehicles
- Generative AI as a creative assistant / super-tool
  - Art, creative writing, music, dance

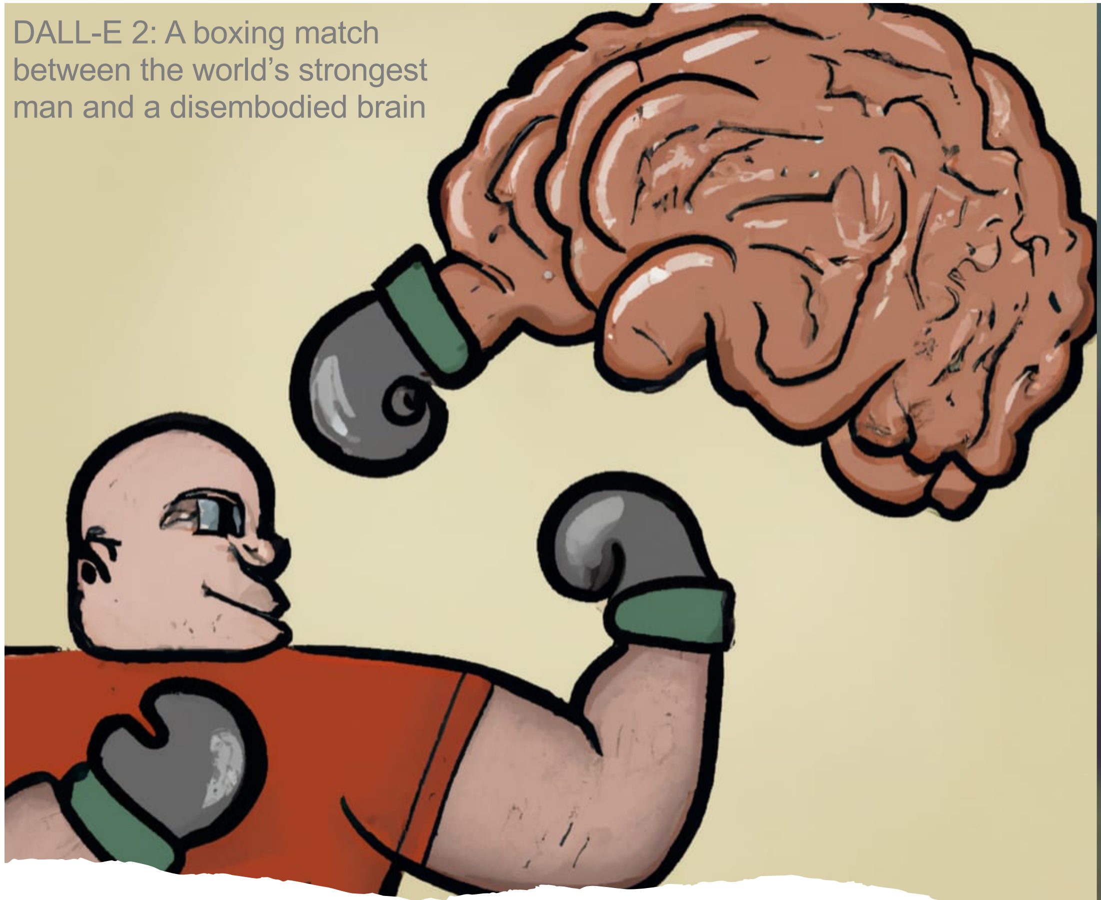
Debates we won’t have time for (until R225)…

- AGI ☞ AGS
- Sentience ☞ Shannon
- Creativity ☞ Pastiche
- Chatbots ☞ Guardrails
- Parrots ☞ Ghost labour
- Alignment ☞ Programming
- Foundation ☞ Language
- Regulation ☞ Monopsony



Moral Codes

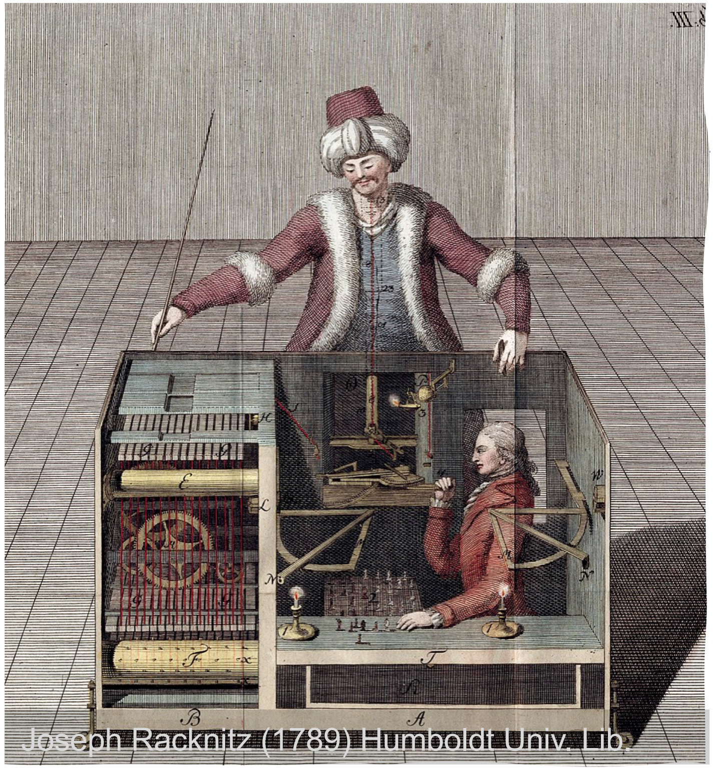
Designing alternatives to AI



DALL-E 2: A boxing match between the world's strongest man and a disembodied brain

SDXL 1.0 A cross between Peter Rabbit and a killer robot from the future

How much "AI" is a branch of literature, not science?



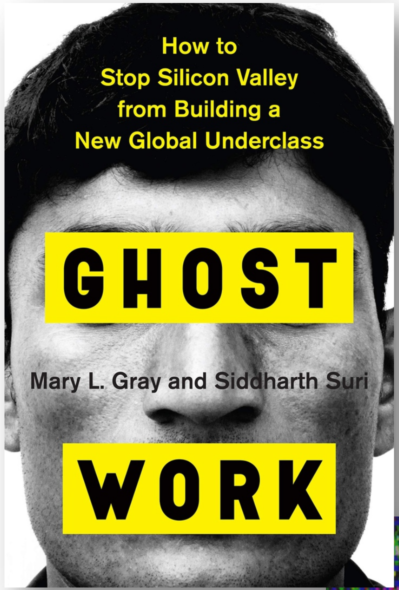
Joseph Racknitz (1789) Humboldt Univ. Lib.

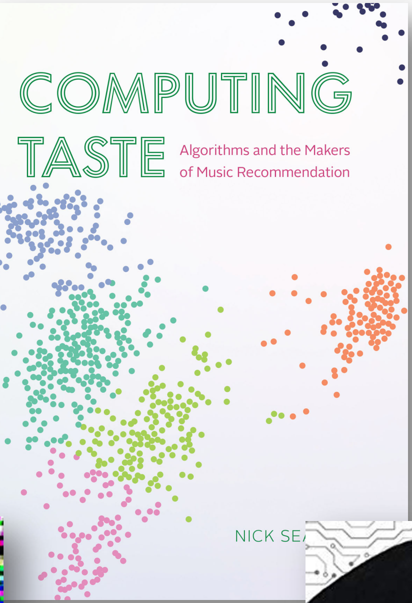
Stable Diffusion: A tarot card reader interpreting a message from the fates

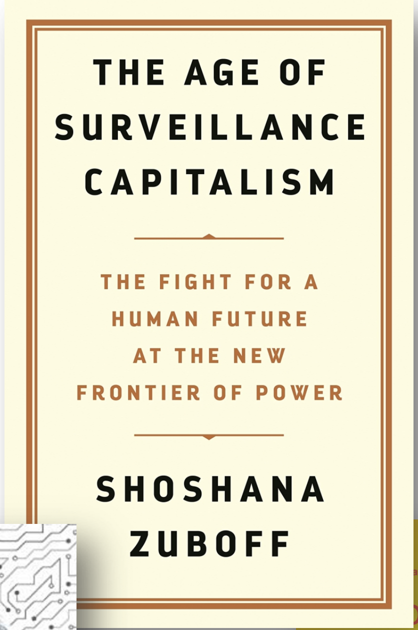
DALL-E 2: A person eating an enormous but comforting lasagne

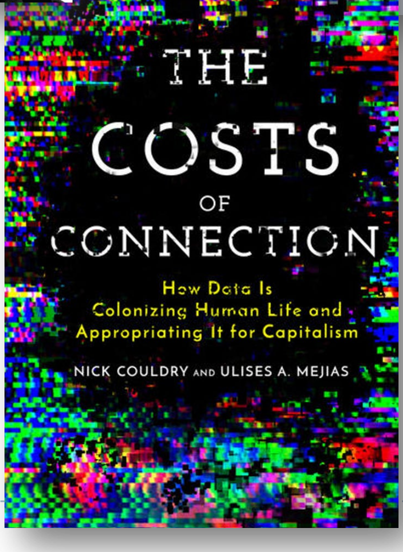
Is ChatGPT …

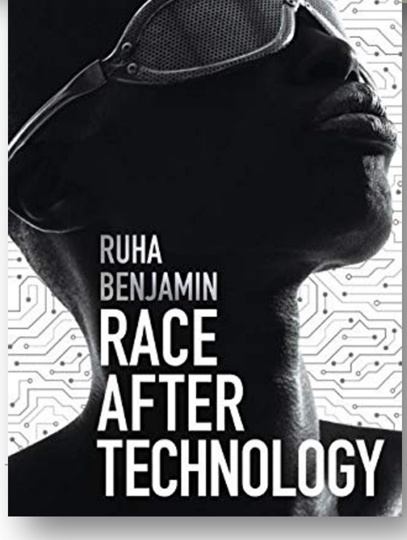
- a Turk (a mechanical stage costume)?
- a Tarot (a shuffled deck for messages from beyond)?
- a Lasagna (a *pasticcio* of comfort food)?

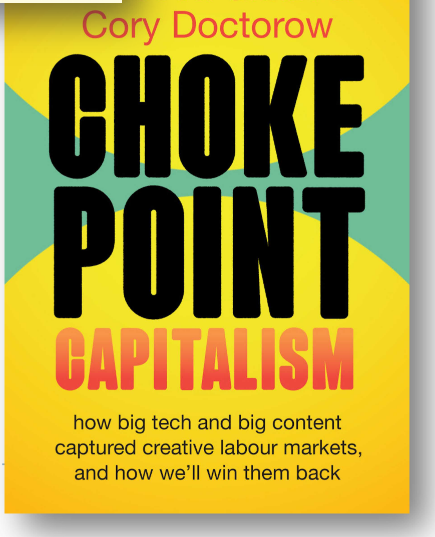

How to
Stop Silicon Valley
from Building a
New Global Underclass
GHOST
Mary L. Gray and Siddharth Suri
WORK


COMPUTING
TASTE
Algorithms and the Makers
of Music Recommendation
NICK SE


THE AGE OF
SURVEILLANCE
CAPITALISM
THE FIGHT FOR A
HUMAN FUTURE
AT THE NEW
FRONTIER OF POWER
SHOSHANA
ZUBOFF


THE
COSTS
OF
CONNECTION
How Data Is
Colonizing Human Life and
Appropriating It for Capitalism
NICK COULDRY AND ULISES A. MEJIAS


RUHA
BENJAMIN
RACE
AFTER
TECHNOLOGY


'An absolute must-read.'
vid Sirota, writer of Don't Look Up
SCRIBE
becca Giblin &
Cory Doctorow
CHOKE
POINT
CAPITALISM
how big tech and big content
captured creative labour markets,
and how we'll win them back

New RSP Unit:
Research Design for Human Participants
Friday 13 October – 9:00 in FW26

The unit will provide an overview of quantitative and qualitative data collection and analysis methods, the construction of open and closed research questions, the design of controlled experiments with human participants, and threats to validity of research results.

Scoping your research
(over to Advait)