Research Study in the Era of AI

Justin Zobel

Pro Vice-Chancellor Graduate & International Research
University of Melbourne

15 April 2024

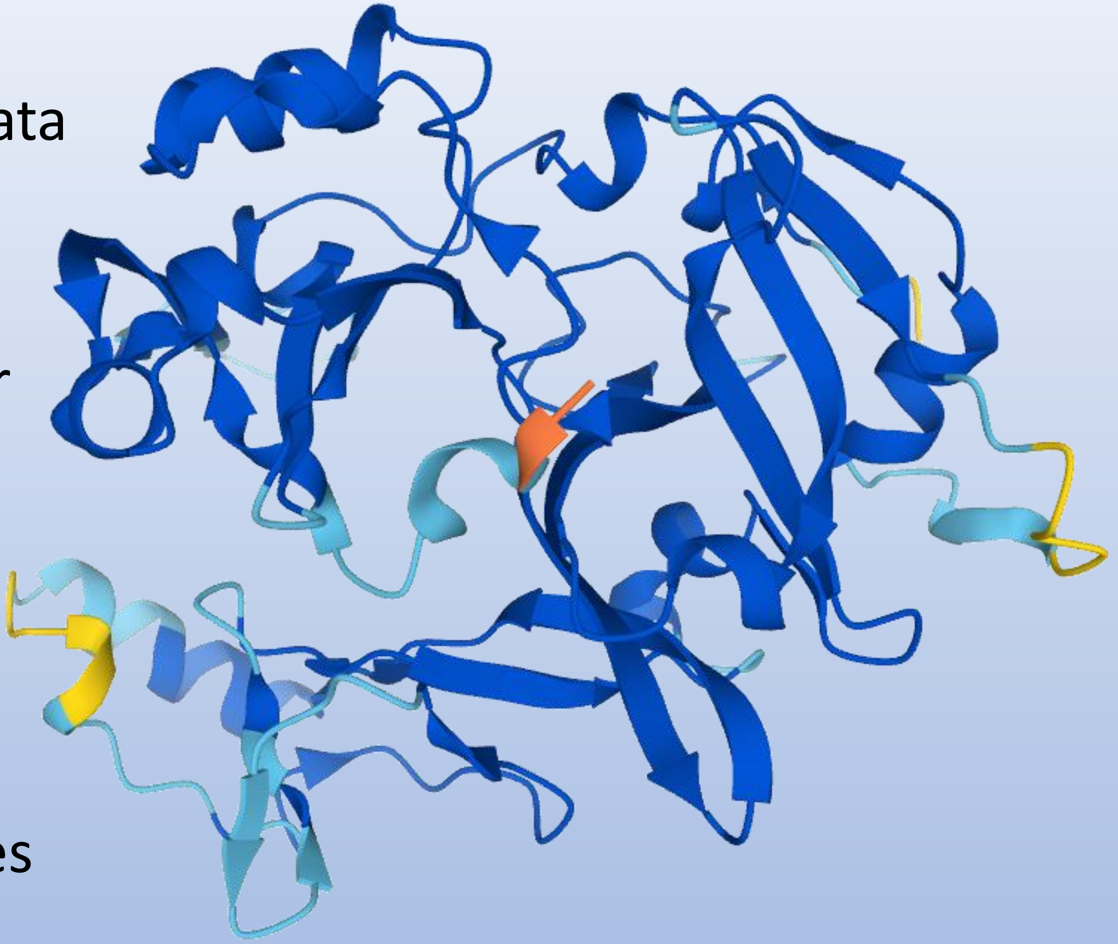
Context: AI technologies in research

They have been widespread for years:

- Discovery of anomaly and patterns in large data sets.
- Disease prediction.
- Optimisation of complex networks (computer chip layout, power distribution, logistics).
- Image analysis.

(These have their own risks but that topic is beyond the scope of this talk.)

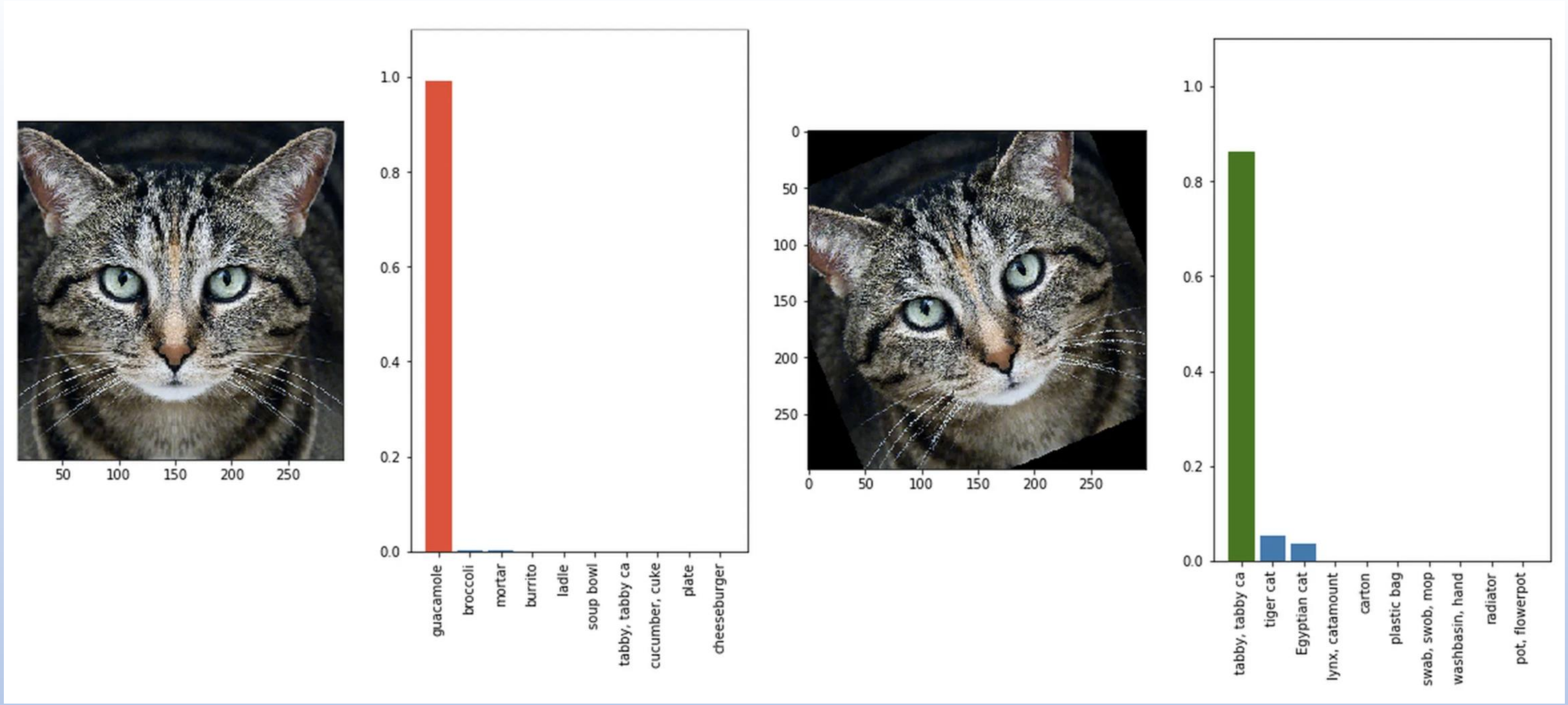
And e.g., *AlphaFold* as a transformative technology in protein analysis – making decades of prior methods more or less obsolete.



https://commons.wikimedia.org/wiki/File:C12orf29_AlphaFold.png

Risks of existing ‘pipeline’ uses of AI

- AI analysis tools are not statistical methods – quantitative measures of uncertainty, error, likelihood of false positives, and so on often don’t apply.
- Layperson tendency to wildly overestimate the capacity and robustness of AI tools (not just in research, but e.g. in legislative processes, medical diagnosis).
- Interpretive mistakes of the kind already common with statistical tools.
  - Lack of quantification of the impact of, or appropriate balance between, false positives and false negatives.
  - Confusion of correlation with true effect (the *ecological fallacy*, the *Lucas critique*); false correlations due to ignorance of real-world characteristics.
  - Amplification of imbalance in input data – that is, they increase bias.
  - Errors and biases due to lack of borderline cases in training data.
- Lack of interpretability of decisions or outcomes; AI mistakes not explicable in real-world terms.



‘The image on the left has been altered so that it’s identified as guacamole. Tilting it slightly means it’s identified once more as a cat.’ James Vincent, The Verge, 2 Nov 2017, https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed

30 November 2022

ChatGPT was released that day; rightly it is seen as a transformative moment in human-computer interaction. It created the impression that it could do *anything*.

? Write about JK Rowling in the style of Derrida [ChatGPT, 26 August 2023 7:51]

> Rowling's tales, while they seemingly coalesce into a harmonious narrative tapestry, harbor an ineluctable trace of irreconcilable difference. The binary oppositions that structure her wizarding world—the Muggle and the magical, the light and the dark—are shown to be ever-porous, their demarcations constantly transgressed by the fluidity of the narrative. This deconstruction of certainties destabilizes the ground beneath our feet, revealing an abyss of signification that refuses to be confined to neat categories.

What LLMs aren’t

- AIs are not models of the real world – they are models of other digital representations of the world.
- They are not a step towards creation of true AI; they are at best a tool that such an AI might use for communication. They don’t *ask*, *know*, or *think*.
  - There is no internal sense of ‘fact’ or ‘belief’.
  - Their answers are not informed by motivation, knowledge, or intention.
  - There is no internal coherence.
- Myths: Garbage out is due to garbage in; there are sparks of intelligence; AI is a substitute for competence; GenAI is a search tool.

‘programs like ChatGPT don’t represent an alien intelligence … [they are] the well-worn digital logic of pattern-matching, pushed to a radically larger scale … what’s been unleashed is more automaton than golem’

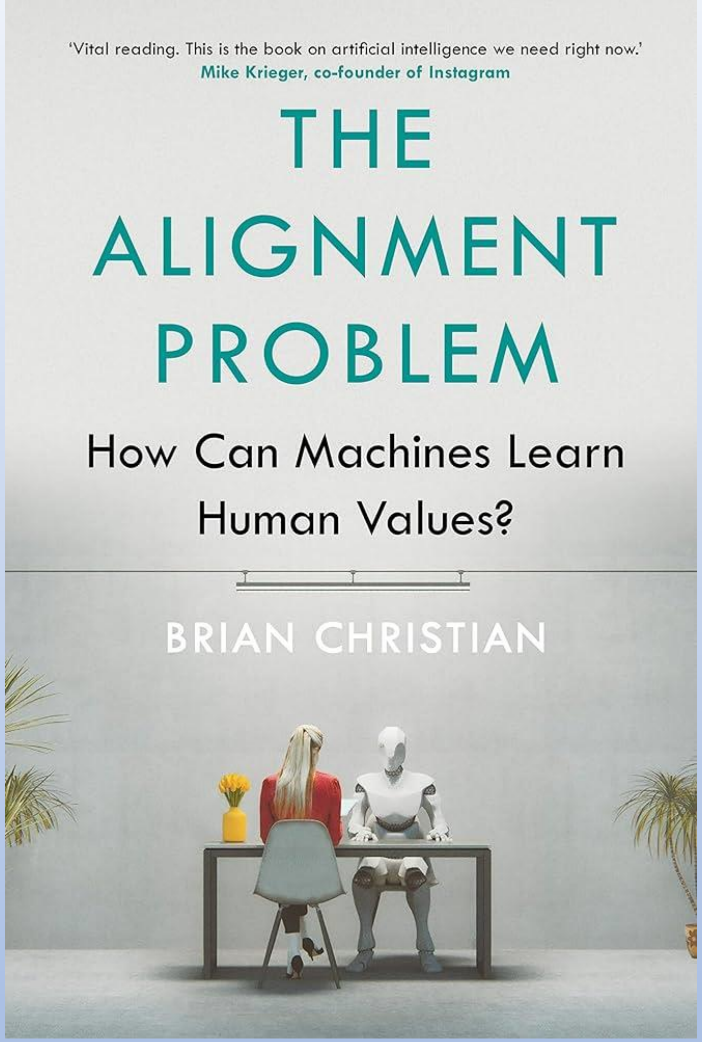
https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have

The alignment problem

- The extent to which an AI is directed towards (intended) human goals, desires, and aspirations.

Challenges:

- There is no technology (existing or in prospect) that can reliably assign a truth value to a statement. Without this, hallucination cannot be eliminated.
- The problem of *external validation* (confirmation, grounding) is unsolved and is not addressed by any recent advances – without which, there is no true AI.



'Vital reading. This is the book on artificial intelligence we need right now.'
Mike Krieger, co-founder of Instagram

THE ALIGNMENT PROBLEM

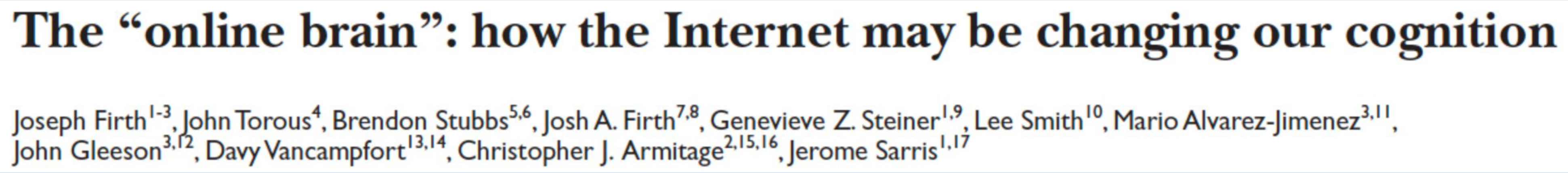
How Can Machines Learn Human Values?

BRIAN CHRISTIAN

Current status?

- Generative AI *is* going to change things – many mundane tasks are being successfully semi-automated (but the market is awash with dubious products).
- Guidelines are proliferating – but they are high-level and not helpful in managing the difficulties, e.g., use of genAI as a ghostwriter. (*So many guidelines!*)
- Students, HDRs, and researchers are making wide use of the technologies.
- Detection doesn’t work and (I assert) will not work in the future.
- GenAI’s limitations haven’t shifted much; understanding of these limitations is improving but there are still many absurd claims.
- It is increasingly embedded in tools, processes, and systems.
- It is only partly monetised and is not good for the planet.
- There are only limited validated case studies of constructive use in research.

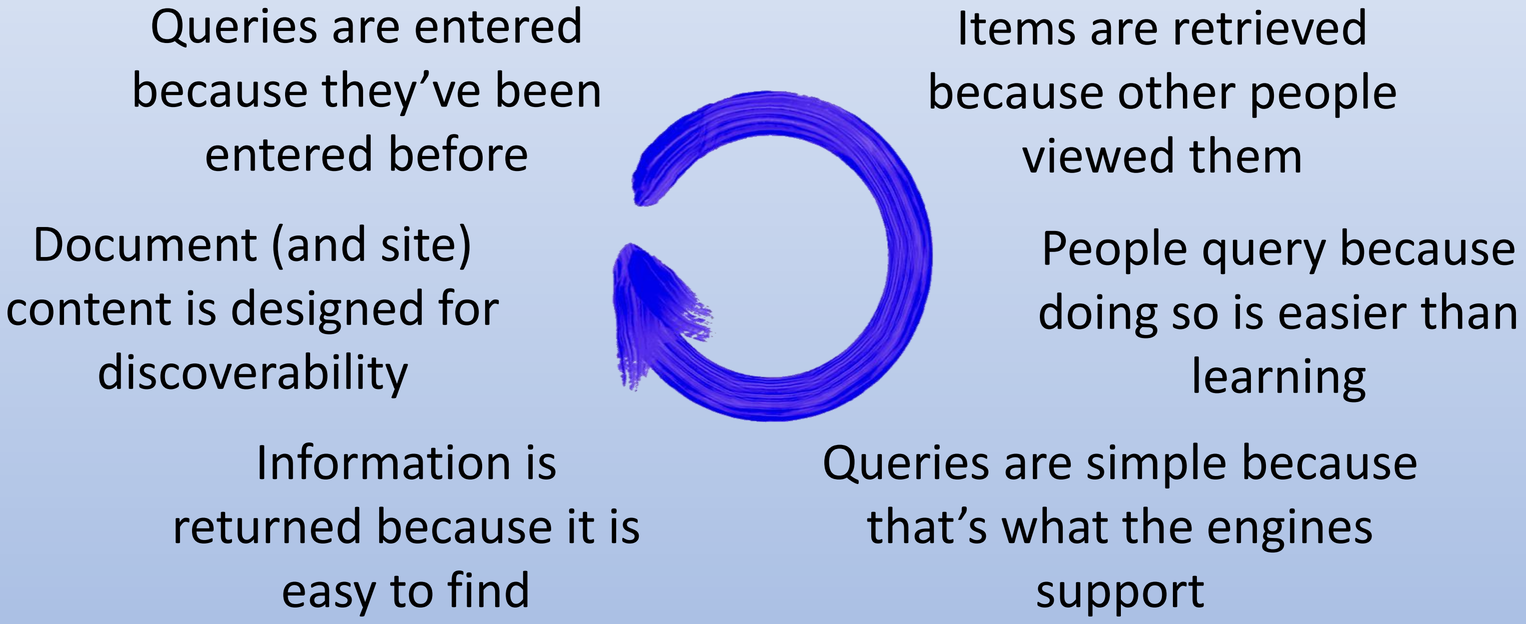
A historical reflection: web search



The "online brain": how the Internet may be changing our cognition

Joseph Firth[1-3], John Torous[4], Brendon Stubbs[5,6], Josh A. Firth[7,8], Genevieve Z. Steiner[1,9], Lee Smith[10], Mario Alvarez-Jimenez[3,11], John Gleeson[3,12], Davy Vancampfort[13,14], Christopher J. Armitage[2,15,16], Jerome Sarris[1,17]

Search is changing knowledge



Queries are entered because they've been entered before

Items are retrieved because other people viewed them

People query because doing so is easier than learning

Queries are simple because that's what the engines support

Information is returned because it is easy to find

Document (and site) content is designed for discoverability

What skills did we preserve or let go as search and online knowledge became pervasive?

Claims about generative AI in research

Allegedly it can be used for

- hypothesis generation
- writing of research proposals
- gathering of literature (i.e., search)
- critique of literature
- experiment design
- simulation of data
- data analysis
- data interpretation
- generation of figures
- every component of write-up
- writing of presentations
- reviewing of manuscripts

… not leaving much left for humans to do.

But some people really, really want it to be able to do the parts of research that they don’t enjoy – and don’t welcome criticisms of, e.g., genAI for lit reviews.

Reflections from a critic

https://helenbeetham.substack.com/p/who-pays-for-authenticity

> ‘Less skilled and productive’ people are being offered the chance to replace more skilled and productive (and presumably more educated) people through the magic of AI. They are not being offered the chance to learn skills, or to earn what skilled people earn, which might be more actually egalitarian. Education is explicitly being replaced by ‘better tech’ as a democratic project.

https://helenbeetham.substack.com/p/deepfake-pedagogy

> Learning is not the accumulation of detail. It is constructing a domain of knowledge … it is, in an important way, more concise than the world it refers to. It is generative … of new responses in new situations. It has [multiple] levels of coherence … It is personal: it becomes part of the self seeing the world, not just the world being seen. This is why we can work out … rules for making our way in the world when we are infants, and don’t then have to boil our heads with data every time we go to sit on a chair.

A specific case: paper reviews, lit reviews?

AIs are (at best) literal summarisers. They can draw on existing critiques but don't have opinions or insight and cannot analyse.

- They are oblivious to characteristics that are key to strong research papers – originality, robustness, correctness, insightfulness of reviews or of interpretations.
- Automatic summaries tend to focus on the elements that authors emphasised – not the elements perceived as valuable by readers.
- They go to the most cited papers, not necessarily the best, most relevant, most reliable, or most useful.

DiffusionBee 2023-08-20 – a photograph of trees in strong sunlight

Why write? Why not use genAI?

Writing *about* something drives *understanding* of that something; *learning* and *retention* are dependent on *effort*.

- Writing – the struggle to state something clearly – is intimately linked with cognition and the ability to organise concepts into a coherent form.
- The act of writing enhances memory during performance of complex tasks.
- Much of research is a fumbling towards ideas and thoughts that have not previously been articulated. The process of grappling with how to precisely express concepts is critical to development of them into research contributions.
- It enables learning of how to explain, articulate, organise, define.
- Assessment and critique of writing is a key tool though which a supervisor can mentor an HDR's intellectual development. Concealment by the HDR of their inability to undertake such writing may block their progress towards completion.

Counterpoint: copiloting is mostly a success

DiffusionBee 2023-07-10 – machine with a lightglobe that represents an idea

Code (software) can be automatically generated and analysed, under experienced guidance.

- This allows a *copiloted* model of code authoring where programmers can be more productive due to rich autocompletion that uses the same methods as generative AI for text.
- Naïve use leads to buggy code – programmers still need to be in the loop.
  - Volumes of code that is faulty are growing much more quickly than in the past.
- Generative AI is an *amplifier of expertise*?

AI literacy in research

Conversations with academics at Unimelb suggest two perspectives on why AI literacy matters.

- First, a concern with the ability for researchers to use (or develop) AI tools in a way that would enable the appropriate production of research.
- Second, a concern with the skills required to discern whether the use of an AI tool helps or hinders the capacity for creative, robust, appropriate, or original research.

Underlying these, academics saw a need for caution:

- Being critical of any information an AI tool produces.
- Understanding that AI won't do the thinking for them.

DALL·E 2023-03-18 05.42.31 - a painting of a professor under attack from robots

This and the following four slides are partly drawn from an internal University of Melbourne discussion paper, 'AI in research and research training: Implications and recommendations' (2024), J Rose and J Brailsford

AI literacy in research …

DeepAI 2024-04-02 - an oil painting of a researcher thinking

Together these mean that a researcher using AI

- Must remain within existing integrity, authorship, privacy, IP, and ethics policies – which, however, will change as AI develops.
- Needs to be able to assess the reliability of different AI tools, their suitability for a given task, and to critically evaluate their output.
- Understands that use of these tools is not a substitute for comprehension or for independent or confirmatory analysis by the researcher.

A hierarchy of technical AI skills in research?

From most to least skilled, the capacity to

- Implement an AI from scratch.
- Create an AI application using an AI toolkit and training data.
- Critique AIs, their basis, and their effectiveness
  - Whether specific data is suitable for training for the intended application – that is, whether the resulting AI technology is likely to be robust on new data.
  - Whether the claims made about the capabilities of an AI technology are sound, e.g., given the experimental evidence.
- Critique claims made about findings that were produced with an AI technology, that is, understand their limitations.
- Understand the principles underlying an AI technology
  - Or understand the metaphors that are used as a sketch of those principles.
  - Be aware of the differences between AI technologies – generative, discovery, inferential, decision-making, …

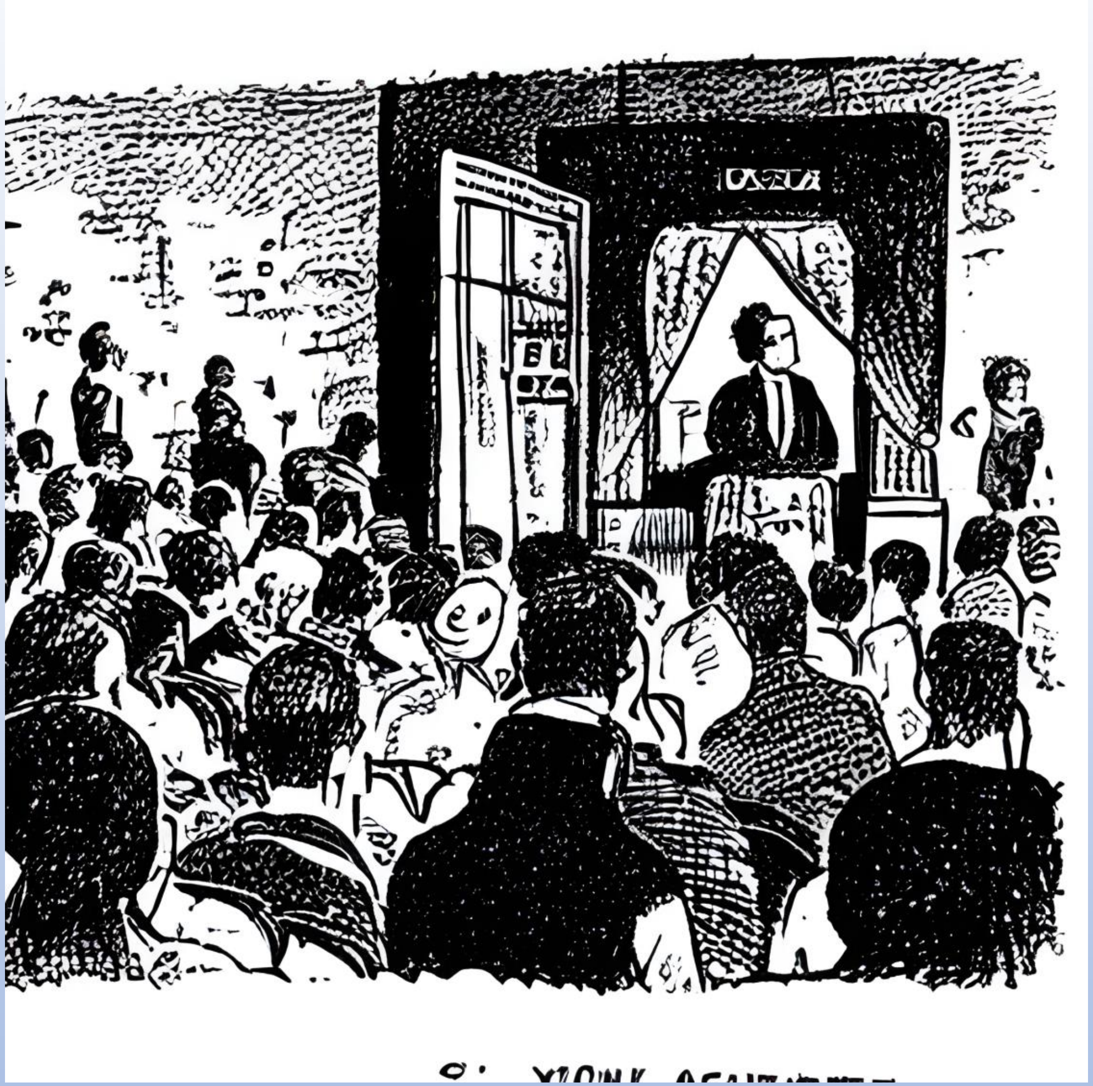
Contextual AI questions for researchers?

There are many ethical & integrity questions in use of AI, e.g.:

- Whether it amplifies or inappropriately propagates bias.
- The extent to which it silently makes use of other people’s IP – e.g., generating ‘novel’ interpretations of an area that are in fact derived from other literature.
- Whether ‘new’ text is in fact plagiarised, or is perceived as the author’s work.

And questions that arise from the limits of the technology, e.g.:

- Whether the extent to which outcomes are categorical is an artefact of using AI.
- The importance of unseen items in training data.
- Whether apparent ease of use has unintentionally limited the scope of a research investigation.
- Whether lack of transparency in how outcomes were derived means that they cannot be meaningfully interrogated.



DeepAI 2024-04-02 - a cartoon showing an audience asking questions

Questions?

Table work #1: AI literacy

1. What don’t you know that you should know?
2. What is the importance of the following dimensions?
   a. Use of AI with ethics and integrity, including understanding of the ethical risks and implications
   b. Evaluation of the quality and suitability of AI for a specific application
   c. Understanding of how the use of an AI might influence the individual’s capacity to undertake research
   d. Understanding of the principles underlying different AI technologies
3. What are the key questions to ask in critical evaluation of (a) AI tools (b) outputs derived from AI tools? How might this compare to, e.g., search tools?
4. How should general AI competencies be developed and evaluated?

Table work #2: Implications for the PhD

1. How might in-progress monitoring or evaluation be changed, and who should be involved?
2. What should be included at submission? What materials should be examined?
3. Is a linear manuscript the best format for evaluating a candidate’s ability to undertake and report independent research? (Counter-example: consider creative PhDs – a work combined with an exegesis.)
4. What elements of pre-AI research skills and capacities do we want to preserve and what can we let go?

Reflections

- We don’t know how much the technology will continue to refine,
  - But the applications are proliferating
  - It’s increasingly embedded in other tools
- Generative AI seems to be leading to magical/wishful thinking and irrational anxieties being presented as fact :

  It seems like generative AI can do anything … wouldn’t it be nice if it could do X … I expect that it will be able to do X very soon now …

  I am so worried about Y and generative AI seems to almost do that … perhaps Y is about to happen …
- Note too ‘the magic of the prototype’ – when something may look nearly finished but none of the difficult aspects have been addressed.
- We need to be wary of complacency and ensure that we’re adapting at all levels – expectations, techniques, behaviours.