Avoiding Sadness

Research policy for Generative AI

Dr Brian Ballsun-Stanton

Introduction

Who am I?

- Dr Brian Ballsun-Stanton
- Solutions Architect (Digigital Humanities) for Faculty of Arts, Macquarie University
- BS and MS in Information Technology, Rochester Institute of Technology
- PhD in Philosophy of Data, UNSW Australia
- A bridge between technologists and academics
- The AI person for the faculty.

Before we get started

- Everyone needs to go register and load up the free version of claude.ai Claude 3.5 Sonnet.
- Load https://perplexity.vercel.app/ (128MB download)
- Play with them during the session, interrupt when you find something interesting or surprising

Motivation

Understanding the capabilities of technology is not a new problem

Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out? (Babbage 1864)

Understanding the capabilities of technology is not a new problem

> *Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out? (Babbage 1864)*

For an LLM: the answer is **YES**.

> *Large language models (LLMs) are fundamentally different from search engines, functioning more as 'vibe-machines' than information retrieval systems. (Ballsun-Stanton and Hipólito 2024)*

Motivation and Aim

- Few folks put in the minimum 10 hours with a frontier model to understand capabilities (Mollick 2024)
- These tools are powerful – do not point at foot, do not pull trigger
- These tools are awful – they are not your plastic pal who is fun to be with.

How do we thread the needle?

This keynote is supposed to be the "positive" keynote.

Basic philosophy of technology

A technology is:

- A tool +
- A technique
- To achieve a desired outcome (Ihde 1979)

Technique matters. No one talks about technique.

Most people don't think about Generative AI correctly

- They think of it as a search engine with access to facts (see ChatGPT is Bullshit (Hicks, Humphries, and Slater 2024))
- They think that it "learns" by chatting (https://simonwillison.net/2024/May/29/training-not-chatting/)
- They see that it responds "I" and think that has meaning.
- They believe the hype
- They have used ChatGPT 3.5 once and gone "that's boring."
- They believe the anti-AI hype

Avoiding Sadness

The title of this talk isn't actually a joke

- My intro to LLM workshop could genuinely have the title "Avoiding Sadness"
- Many people have incorrect assumptions of what these tools are and do
- Sadness, here, is the mismatch between reality and our expectations. It is likely, without thought and attention to detail.

Macquarie's guidance note

https://policies.mq.edu.au/download.php?associated=1&id=768&version=1

- Researchers are allowed to use it
- Supervisors must closely and actively supervise their postgraduate students when their students use it
- Acknowledgement of use (BUT NEVER CITATIONS) is required
- There are narrow places where it is forbidden, for cause

Allowed use

At the end of the day, when a researcher puts their name to their work, they own the consequences.

- These tools *manipulate* words. They are not search engines.
- The manipulation of words and the production of "vibes" is shockingly useful.
- They need to consider which tools they use and how they use them.
- Researchers own the consequences of their judgement. **Nothing has changed.**

Postgraduate use

Ultimately, it is the supervisor's fault if their student uses a generative AI in an incorrect way.

- Active supervision is just that: the student needs to engage within the guardrails and feedback of the supervisor
- The supervisor cannot disclaim responsibility
- If the student ignores the guardrails, it's just like any other failure of candidacy.

Bad outcomes

- It is not an author, it cannot be fired. Therefore, it cannot be cited or credited as a co-author. (It cannot take the blame.)
- These tools are not presently capable of judgement or working with tacit knowledge. Using them for peer review or originating ethics applications are forbidden.
- Cannot be used to wholly generate works and claim them for yourself.

Prepared questions from the audience

This is a triumph...

Re: LLMs represent "a triumph for the humanities" - can the humanities as they are now be said to triumph, or do we need to include much more knowledge about LLMs into humanities curricula in order to increase understanding of their 'grammars' before they can "triumph"? How would you address such changes within the curriculum?

This is a triumph...

- Suddenly, how you say something matters.
- Knowing something about something and asking the right questions will be the fundamental skills of the age (Vinge 2006)
- Our present textual skills will serve us well: word use, context, source evaluation, critical thinking
- This is an *opportunity* for triumph for the humanities at the same time as it's an crisis for universities.
  - Knowledge demonstration has been alienated from the student
  - The value of the university is in question
  - Employers are mistaking these things as opportunities for automation, rather than mechanisms for worker augmentation.
- The dreaded word: workload.
- Sample undergraduate unit policy

Invocations are magic

What kind of knowledge system can we create around, how can we operate with the concept of "knowledge" around tools that are so powerful and so little understood? As you suggested with the term "grimoire", for many people this will remain a little bit like "magic" - what does that do to our present understanding of knowledge and a knowable, scientifically understandable world that we are competent actors in?

Invocations are magic

- They're a year old. Give me a break.
- We have a duty to educate, but also we – as those in the humanities – have the fundamental skills of use.
- Many people have a shallow understanding of computing. This is just one more opaque nail in the coffin.
- They manipulate text. They have no relationship to knowledge at all. Tool + technique. We are responsible for technique and objective.

Power imbalances

Should – and how could – universities address power imbalances / knowledge asymmetries within LLMs? E. g. the majority of training data being in English / from US/Anglo perspectives, even if the service is being used in a different language?

Power imbalances

- This question is a good question, because the assumptions illustrate a fundamentally wrong conception: knowledge-in-model. This is a trap.
- We should not treat models as knowledge-having things. They are word-manipulating things.
- The power imbalance is one of equity and access, not of training sets. (WolframRavenwolf shows German testing protocols)
- Corpus size matters, and the data is not english-only data. Larger models are better, even if the smaller german-only models do exist on huggingface.

What's the point of uni?

Do you agree that the main role of universities is to teach students how to think critically? In this regard, how should AI technologies be incorporated into the study process so that they help rather than create obstacles when it comes to achieving this objective?

What's the point of uni?

- Begging the question.
- We have multiple objectives. One of them will be to have students be able to judge *when* to use a tool, which tool to use, and to be proficient in its use.
- Generative AI is a tool.
- This also demonstrates the central catastrophe awaiting us.

How do we produce new knowledge?

How will AI generation impact human ability to produce new knowledge, particularly in social sciences?

- It's a tool. Powerful, dangerous, easily misunderstood, poorly used.
- We're looking at the edge of Kuhnian normal science.
- Emperor has no clothes situation.

A contradiction of use

You warned not to use LLMs for value judgements or ethics questions – but you have also used it to help with e. g. preparing ethics reviews. How do you resolve this?

A contradiction of use

- (Load actual use, in Claude)
- The *manipulation* of text for plain text wording is not the same as the *creation* of text.
- It's great at rephrasing, unpacking, and framing tacit knowledge.
- It just shouldn't be used to create, whole-cloth.

Data and privacy

*Data privacy implications - what parts of research can these tools be used for without SERIOUS data issues? Are there any models/terms of service that can be safely used within the research process? * Examples: if I want to use them to line edit an unpublished paper that includes research data from a collaborative project, do i need to ask everyone on the project for permission? * Which models will not store and henceforth own that unpublished data that I have given them?*

- Most, actually. It's just in the terms of service
- Prefer API use to chatbot mode
- Read the privacy policy and pay for them
- Absolutely get permission from colleagues, just as you would any other research method or tool use.
- Anthropic is quite good around its (paid) terms of service.

References

Babbage, Charles. 1864. "Passages from the Life of a Philosopher." https://www.gutenberg.org/files/57532/57532-h/57532-h.htm.

Ballsun-Stanton, Brian, and Inês Hipólito. 2024. "Is the 'Calculator for Words' analogy useful for communicating about LLMs?" Zenodo. https://doi.org/10.5281/zenodo.12602858.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Ihde, Don. 1979. "Heidegger's Philosophy of Technology." In *Technics and Praxis*, edited by Don Ihde, 103–29. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9900-8_9.

Mollick, Ethan. 2024. *Co-Intelligence: Living and Working with AI*. New York, New York: Portfolio/Penguin.

Vinge, Vernor. 2006. *Rainbows End*. 1st ed. New York: Tor.