Ethical Use of Artificial Intelligence

https://bit.ly/4df8fAY

Dr. Ben Walter

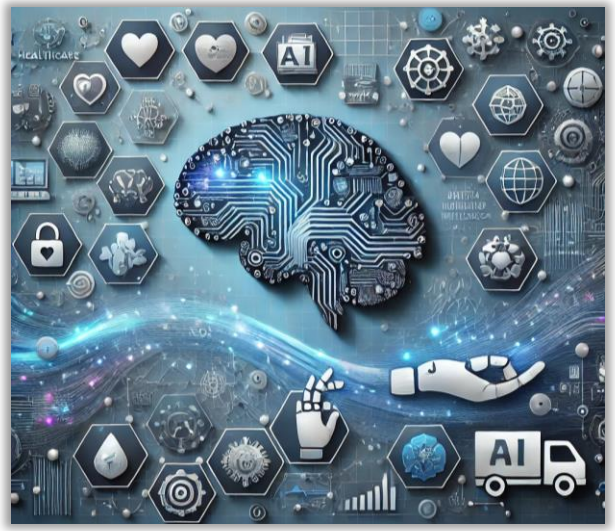
Emerging Caribbean Scientists Seminar

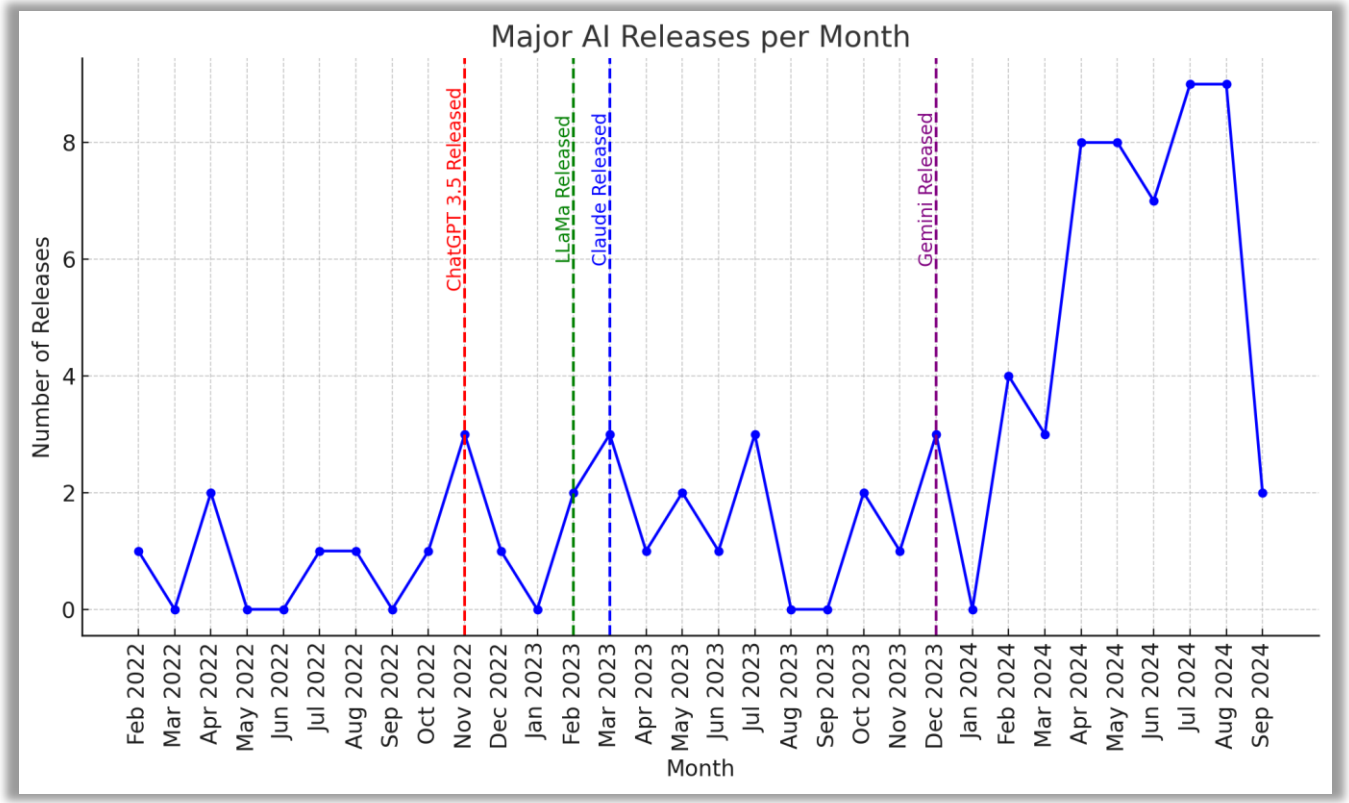
September 25, 2024

Artificial Intelligence (AI)

AI is “any computer system or application that performs tasks that normally require human intelligence, such as perception, reasoning, learning, decision making, or natural language processing.”

-- Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Oct 30, 2023





Data from https://nhlocal.github.io/AiTimeline/

Alan D. Thompson (https://lifearchitect.ai/) estimates that ChatGPT currently outputs the equivalent of *the entire printed works* of mankind (130 trillion books averaging 70k words per book)[1] every **two weeks**.

[1]Based on Google Books study

Major AI Platforms *(ranking as of Sept. 20, 2024)*

ChatGPT (OpenAI)

May 2024 **GPT-4o** - multimodal input / output

Sept 2024 **o1** - multi-step reasoning
- advanced math / physics

Gemini (Google)

July 2024 *AlphaProof* scores silver rank in IMO

Aug 2024 **1.5 Flash** - targets research
- includes some fact checking
- includes some references (partner with OpenStax)
- can retrieve data from web

Claude (Anthropic)

June 2024 **3.5 Sonnet** - focus on ethics, alignment, safety
- "constitutional AI" model
- accuracy over creativity
- does not train on user interactions
- artifacts!

Grok (xAI)

Aug 2024 **Grok 2** - "sense of humor"
- available on X

LLaMA (Meta)

Open source, freely available

July 2024 **LLaMa 3.1** – in Facebook, Messenger, Instagram, WhatsApp

LE CHAT MISTRAL (Mistral AI)

Emphasis on free / open-source models

July 2024 **Mistral Large** / **Codestral** / **Mathstral**

perplexity (Perplexity AI)

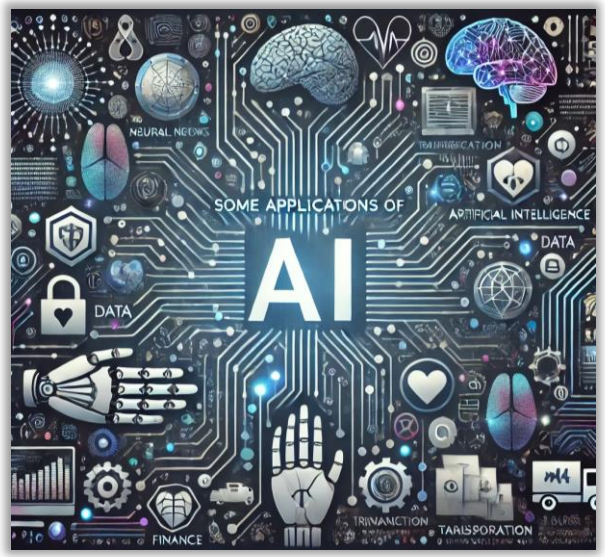
"AI search engine" (LLaMa backend)
- searches internet for answers
- summarizes results and gives reference links

Copilot (Microsoft)

- based off GPT-4o
- available in all Microsoft products (extra cost)

Some applications of artificial intelligence

- Self-driving cars
- Smart home devices
- **Virtual assistants**
- Fraud detection
- Chemical research
- **Customer service**
- Weather forecasting
- Market prediction
- **Recommendation systems**
- Facial / object recognition
- Sentiment analysis
- **Speech / language translation**
- Voice-to-text (meeting / video captions)



- **Content summary**
- **Content editing**
  (text, computer code)
- **Content generation**
  (text, images, video, audio, speech, code)
- Content moderation
  (message boards, chats, online game interactions)
- Personalized learning / tutoring
- Help with instructor grading / feedback
- **Chatbot**
- Spam filtering
- Network intrusion / virus detection
- Sportscasting / commentating (e.g. Wimbledon)

Large language model (LLM) AI's learn patterns in human created texts
- Trained on extremely large datasets covering wide array of topics
- Responses appear human-like, comprehensive, and researched
- Can answer specific questions, generate ideas, summarize, translate, etc.

Some problems with artificial intelligences (LLM)

Issues with accuracy (though rapidly improving)

- "Hallucination"
- Lack / fabrication of references
- Biases / gaps in training data

"Poisoning the well"

- Proliferation of AI data corrupts future training
- Causes decline in *quantity* and *quality* of human-generated content *(and loss of expertise?)*

  → artists, writers, programmers (*stackoverflow*)

Manipulation / Misinformation

- Deep-fakes and the liar's dividend
- "Algorithmic radicalization" / "Rabbit-Hole effect" (via maximizing engagement)
- Intentional misuse; see *Costello et. al*. "Debunkbot" (2024)
- Hacking weights and features; see *Templeton. Scaling monosemanticity… (2024)* and *Zehavi. Facial misrecognition systems… (2023)*

Ethical Considerations

- Bias and discrimination
- Fair access and inclusivity
- Privacy and surveillance
- Transparency and explainability
- Accountability and accuracy
- Authorship rights
- Plagiarism and appropriate use

Bias and Discrimination

AI reflect any biases present in the data they are trained on, perpetuating existing inequalities. *(Generated content is based on learned patterns.)*

- **Bias in predictive policing / hiring / automatic decision making**
  Predictive algorithms trained on historic data (e.g. historic crime data or cv's of successful applicants) don't account for changing demographics.
- **Bias in underrepresented language / dialect queries and responses**
  Low prestige languages and dialects receive less informative responses. This perpetuates sociolinguistic inequality.
- **Discriminatory results from machine learning algorithms due to training data bias**
  E.g. facial recognition prone to errors on darker skin tones yielding discriminatory outcomes; medical algorithms less effective on underrepresented groups yielding poor diagnoses; image generation reflecting training bias.
- **AI grading / plagiarism detection / peer review automation**
  May have bias against certain styles of writing (especially from non-English-speaking backgrounds) or may flag legitimate work as plagiarism.

Fair access and inclusivity

The gap between those with access to AI and those without can widen existing social and economic inequalities.

- **AI cost**
  The cost of access to high quality AI further segregates economic classes.
- **Disparities in access to AI education**
  Underrepresented groups may have fewer opportunities to learn AI skills, preventing them from participating in or benefiting from the “AI economy”.
- **Accessibility to AI enhanced learning**
  Changes are promised by personalized learning / tutoring platforms that adapt content to individual students’ needs (for example “ChatGPT Edu” / “MathGPT” / Gemini’s partnership with OpenStax).
  Consistent, *safe*, and ***secure*** access may only be available to few languages / demographics / socioeconomic classes.

Example: Harvard gives ***all*** students access to ChatGPT Edu along with custom built sandbox of isolated AI’s.

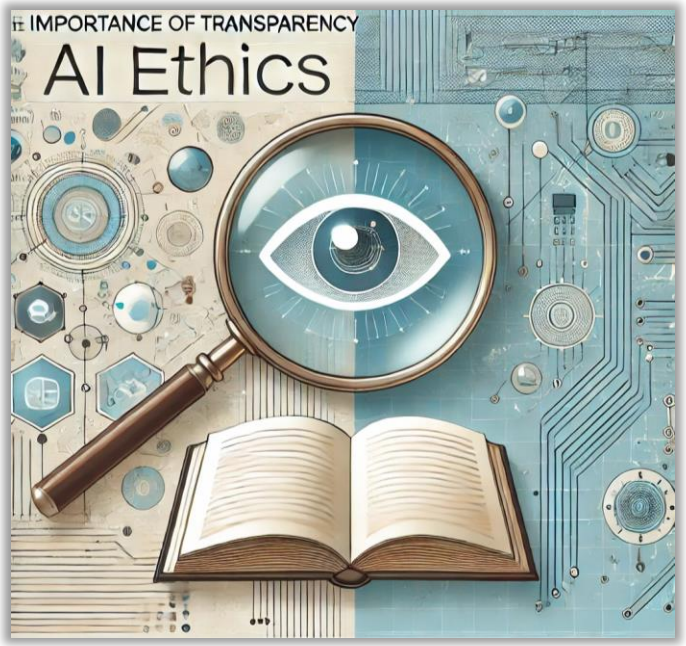
Privacy and surveillance

AI can be induced to leak training data. Massive training data sets can also be directly stolen or exposed via human intervention or mistakes.

- **Training data privacy**
  Internal AI may be trained with confidential data; e.g. names, phone numbers, addresses, salary. Known attacks can extract this information.
  - see *Nasr. Scalable extraction of training data… (2023)*
- **Query data privacy**
  Many AI will *self-improve*, training on supplied query or analysis data. Any non-anonymized data exposed to the AI is at risk of leakage. (*Read AI???)*
- **Social media scraping privacy**
  See *Cambridge Analytica scandal* from 2018 (harvesting personal data from millions of users to influence politics).
- **Overcollection of data**
  AI systems tend to collect extra data to maximize effectiveness; e.g. home assistant recording ***all*** video and audio.
- **Privacy of children**
  See FTC lawsuit vs ByteDance due to collecting and hoarding data on minors (and impeding parental access).

Transparency and explainability

AI and machine learning systems operate as a "*black box*". Training generates numeric *weights* for *features* – *"thought process"* / reason for decisions is not always clear. Necessity for "human in the loop".

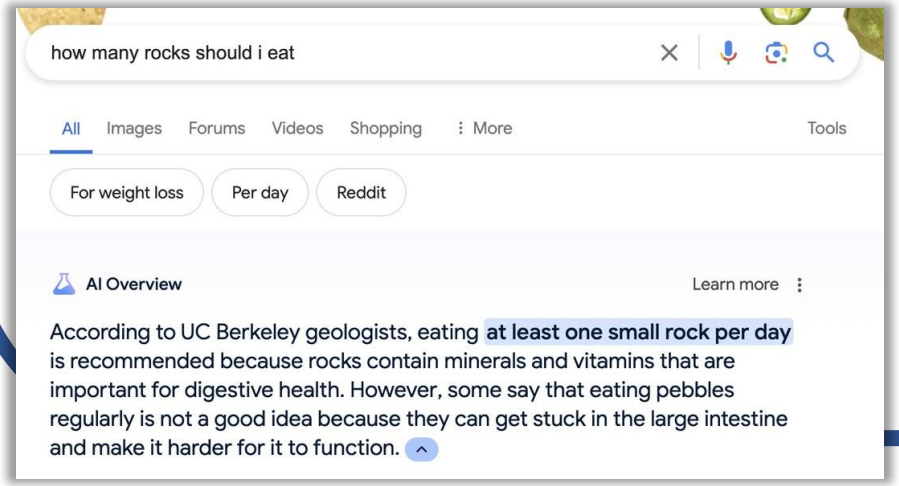


- **Decisions sometimes based on unrelated features**
  In early radiology application, AI actually based decision on **type** of scan.
- **"Algorithmic collusion"**
  ML driven high frequency trading may contribute to market volatility (2010). AI driven house price suggestions accused of price-fixing.
- **Inscrutable decisions have major effects**
  AI used for administrative purposes (e.g. university predicting student success, identifying at-risk students; credit card deciding credit worthiness; bank deciding interest rate on loan) can have major affect.
- **People may not even be aware AI / ML is involved**
- **AI training data may be shared / sold without user knowledge**

Accountability and accuracy

AI make mistakes, but take no responsibility. It is the user's responsibility to check for errors; though there are some grey areas…

- **Responsibility for self driving car accidents?**
- **"Blackout challenge" lawsuit vs TikTok?**
  Aug 28. US appeals court revives lawsuit: ML recommendation engines "not protected by sec 230 of Communications Decency Act"
- **National security threat order (PAFACAA) vs ByteDance?**
  TikTok collects data on US citizens, could be used to manipulate opinion?



• Output may not always be accurate. You should not rely on Output from our Services as a sole source of truth or factual information, or as a substitute for professional advice.
• You must evaluate Output for accuracy and appropriateness for your use case, including using human review as appropriate, before using or sharing Output from the Services.
[…..]
• Our Services may provide incomplete, incorrect, or offensive Output that does not represent OpenAI's views. If Output references any third party products or services, it doesn't mean the third party endorses or is affiliated with OpenAI.

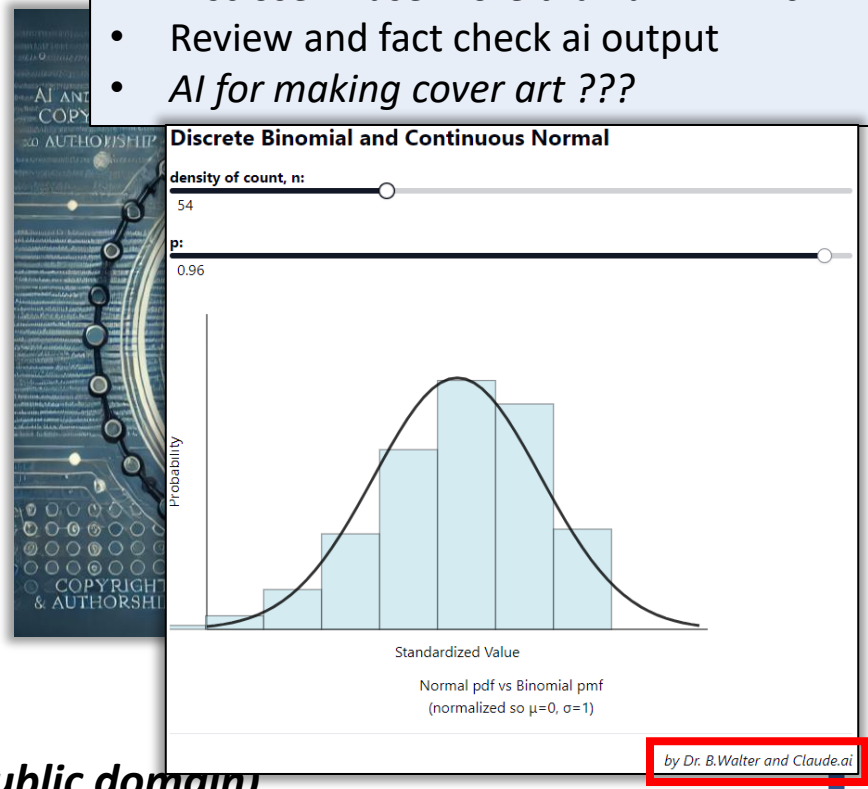
- OpenAi terms of use

Authorship rights

Best practices from authorsguild.org:

- Use AI for support, not replacement
- Rewrite any AI generated text
- Disclose AI use more than *di minimis*
- Review and fact check ai output
- *AI for making cover art ???*

AI trained on data from across internet and other sources. *Frequently violating copyright.* Generated data can also violate copyright accidentally or intentionally *("…in the style of…")*.
There are also issues about ownership of generated content.



- **Fair use and training data**
  Multiple lawsuits vs OpenAI about scraping images, books, YouTube
- **Accidental plagiarism**
  See Scarlett Johansson voice dispute with OpenAI
- **Credit if AI plays substantial role in research / writing**?
- **US copyright office gives NO ownership to any AI created content *(all is public domain)***
  See "Monkey Selfie" lawsuit and Zarya of the Dawn comic dispute.

Ownership of content. As between you and OpenAI, and to the extent permitted by applicable law, you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output.

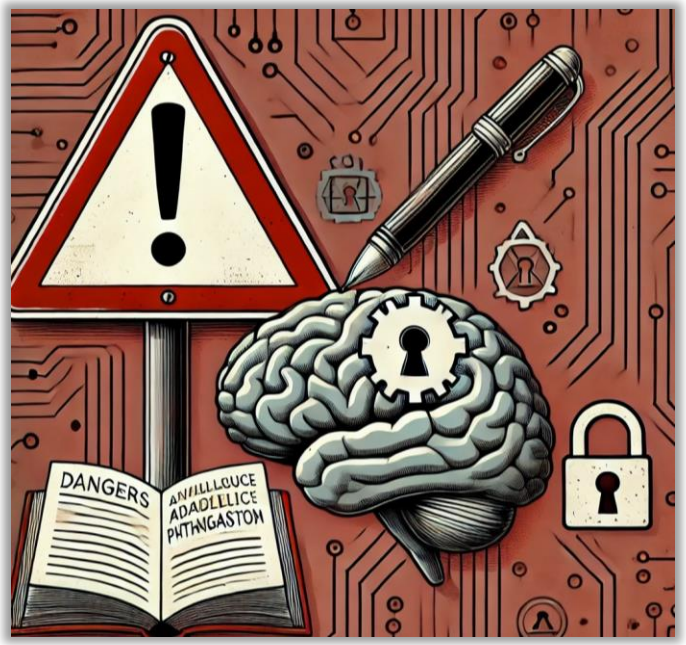
Similarity of content. Due to the nature of our Services and artificial intelligence generally, output may not be unique and other users may receive similar output from our Services. Our assignment above does not extend to other users' output or any Third Party Output.

- OpenAI terms of use

Plagiarism and appropriate use

When used as a *replacement* rather than *augmentation* for humans, AI can have long-term detrimental effects. Also issues with malicious forgery.

- **Erosion of research / critical thinking skills and creativity?**
  Misuse of AI can undermine the skill development goal of assignments.
- **Academic integrity?**
  Currently impossible to reliably identify AI.
- **Forgeries.**
  Range from *"… in the style of…"* prompts to "deep-fakes".
- **Illegal / objectional content generation.**



"Human in the loop"

- Automated ML decisions are dangerous!
- Important for responsibility and accountability
- Don't blindly copy / believe AI output!

Don't enter confidential data into 3rd party AI

Use AI to augment your creativity, not replace.

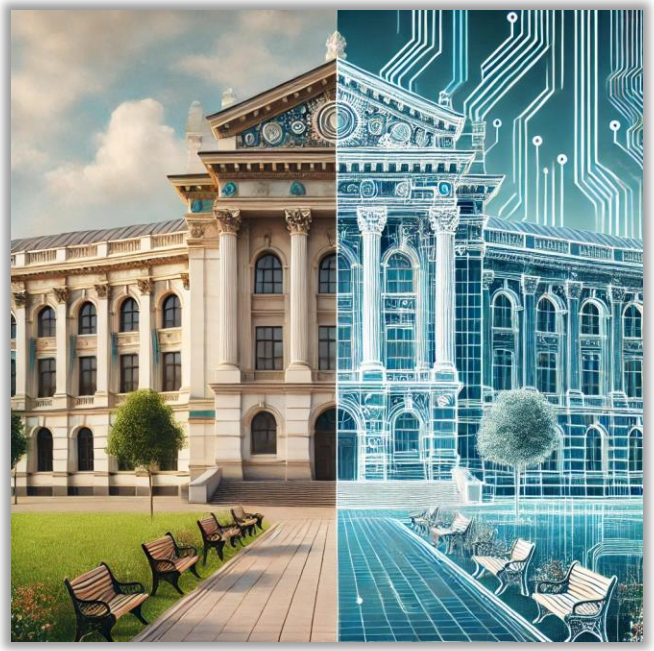
Think critically about likely training data and bias.

Policies

IEEE Ethically Aligned Design

120 AI bills currently in Congress.

EU AI Act. (Jun 2023)



Harvard guidelines for students:

- Protect confidential data
- Students are responsible for content they produce / publish
  - Review for accuracy / copyright infringement
- Students must cite their use of AI
- Students must adhere to course policy
- Be alert for deep-fake phishing

University course policies (for individual classes):

- AI use prohibited
- AI use allowed with permission
- AI use allowed with acknowledgement
- AI use allowed freely

Some references

https://bit.ly/4df8fAY

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814.

Templeton, A. (2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.
https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Zehavi, I., & Shamir, A. (2023). Facial misrecognition systems: Simple weight manipulations force dnns to err only on specific persons. *arXiv preprint arXiv:2301.03118*.