



ART Submission to Supporting Responsible AI consultation

Closing date 26 July 2023.

Summary of recommendations

1. Datasets

Data must be of good quality to stop generative AI from working badly. It must be clean, complete and align well with the complex evolving real world situations that generative AI might meet

The risk is that there is **no set of standard rules** that apply to the whole of the data content that may be used by AI developers to train and provide content to generative AI's

What is needed is **a data commons, open to all, maintained by all, but ring fenced by a mandate that it is the only legitimate site for providing clean reliable ethical data** to be used by AI developers.

The data commons could **be owned by an independent supra national body** on analogy with national ownership of crown land. That body could be charged with determining who was able to use the common resource, and under what conditions.

To regain access to top quality data, **a major revision should take place on the use of copyright and intellectual property laws to privatise what should be publically available peer reviewed academic knowledge** produced by universities and research institutes.

An International monitoring group modelled after the Intergovernmental Panel on Climate Change (IPPC) should be set up to provide comprehensive assessment reports about the state of scientific, technical and socio-economic knowledge about AI and its risks.

- It should be tasked to **assess the risks inherent in the commons data pool used for training AI** in concert with a number of other monitoring and reporting tasks.
- It should have **oversight of the quality of the data commons**.

2. Ethics in the machine

AI models will have value systems, whether intentional or unintentional.

"Constitutional AI" models provide an opening for humans to steer AI value systems relatively easily.

It is particularly useful for ;

- Providing **a solution to "scalable oversight"**,
- **Creating decision transparency**,
- it enables **easy adoption of new ethical principles** by all AI producers
- It makes Constitutional AI much **simpler for monitoring organisations and regulators to oversee**.

- **Parameters can be set for acceptable behaviour through principles guided by regulators.**

The ultimate value of testing a **constitutional approach to developing generative AI models** however is that it **can be adopted by companies and other organizations and spark further interest in how they might design and adapt ethical AI constitutions.**

However protections should still be provided against harms not controlled by constitutionality.

It would be wise to;

- **prevent inversion of ethical constitutions for misuse**¹
- **not use Constitutional or any other AI for automated decision-making.**

With respect to the problem of ethical inversions of AI constitutions, **controlling the building blocks of ethical AI such as access to clean comprehensive massive datasets and limiting the use of them to licensed users** might help to prevent inversion of the Constitutional model for unethical purposes

Where automated decision-making is implemented:

- There should be a **clear path for those affected by decisions to seek review**
- A **central registry of all use cases for AI should be kept** which clearly sets out the use case itself for each instance, and which explains in plain language how the AI in question would work. The ease of publication of the operating principles of Constitutional AI mean that **real time publication** at the time of development in a public registry is both desirable and feasible.
- Business rules and algorithms should be made available, to enable independent expert scrutiny.

3.Users and use

3.1 Risky Users

A **unique identifier** on content produced by AI such as a watermark or pixel level identifier would help to verify "good" AI, but it would do nothing to help to unearth "bad" AI.

The EU is drafting **rules to ensure that artificial intelligence is trustworthy and safe** and is developed and used in compliance with fundamental rights obligations. Australia should follow suit.

3.2 Risky Use

A short list of unethical uses of AIs to be prevented as a matter of urgency includes;

- Activities that cause **psychological disruption or destruction**; e.g., attempts at mass manipulation of social or political beliefs and sentiment especially for political, but also for commercial ends -
These include fostering social decay, ostracism of specific groups, destruction of social support networks etc.
- Activities that cause **physical disruption or destruction**; e.g. war, famine, disease, and damage to the biosphere.
- Activities that **infringe on personal rights**; e.g. real time facial identification, or use of personal data without permissions and protections.

¹ See p20 this essay and p11 Ganguli et al 2023

- Activities that might cause **economic harm** to states or individuals
- Activities that **might prevent the discovery of systemic harms** such as those listed above which originate from other causes. ('Whistle blowing by AI'.)

4. Changing the system incentives.

4.1 Regulation for accountability

The whole AI ecosystem, including the datasets mentioned above, suffer from lack of **set of standard rules** that could govern it.

Provide producers with a **legal identity as an AI producer**. This is necessary to bring a disparate set of players under one identifiable heading and to locate them to ensure compliance.

Licence producers and users on analogy with producers and users of other dangerous products, eg guns

An AI licensing system might have the following features;

- In order to be legal, AI producers and users should be licensed.
- A mandatory requirement of any proposal to design an AI for a general or a particular use should be to provide as a condition of licence, a transparent use case which is open to the public to examine and to critique for both intended and unintended effects before the AI is built. The use case should conform to a set of mandatory ethical guidelines.
- Operational AI models should be required to demonstrate that they operate in accordance with the design principles they submitted in order to gain clearance to build.
- Legacy AI should also be required to demonstrate conformity with the mandatory ethical design principles set as licensing standards.
- Licensing standards should also address product lifecycle control, covering all phases of conception and purpose, development, use and monitoring and evaluation.
- Access to the aforementioned "data -commons- with- a- fence" would be conditional on being licensed. This is a way to ensure that at least one contributor to unethical outcomes, bad or missing datasets, is vetted for quality. Guaranteed access to a superior source of datasets is also an incentive to become a licensed producer.
- In addition, some forms of currently available alternate datasets should be illegal to use, especially those with high levels of disinformation and biases.
- Both kinds of licensees - producer and user- should be regularly checked for compliance to licensing standards.
- For producers, monitoring and evaluation of compliance and of new evolutions of models on a regular basis should be built in to the conditions for continuing to hold a license.
- Lastly, licensing might give preference to organisations structured to reduce the perverse incentives of hyper competition, for example not for profit entities and co-ops with a broad range of stakeholders. Preference might also be given to organisations that behaved as if they were offering a public service in the public interest - eg those with altruistic motives, those that offer free products usable free in perpetuity.

4.2 Regulation to correct social misalignment of incentives.

- Design legal frameworks, policies and institutional design with a view to distributing rather than concentrating power. State capture is a real possibility. AI ownership itself, without considering any other effect, can confer almost despotic levels of power on its owner, be they human or the machines themselves.
- Focus on public interest projects and organisational architecture and build the concept of the fiduciary duty of AI producers to act in the best interests of the public good.
- Globalise AI governance through independent agencies to the point where no bad actor is immune from discovery and sanction, but foster communitarianism, care and mutual responsibility at all levels of governance to foster a culture of responsibility rather than one dominated by competition. Only through **global cooperation global monitoring and global controls** can bad actors be subject to regulation that might especially **prevent the deliberate development of AIs to supersede humans**. such objectives might be under the umbrella of something like an AI equivalent of a nuclear non-proliferation agreement. A global compact also may enable a sufficient extent of monitoring to **prevent the spontaneous development of such capacities**.
- Find ways to foster altruism and altruistic AI providers.

ART Submission to Supporting Responsible AI consultation

This submission is made by Dr Julia Thornton on behalf of Accountability Round Table 26/07/2023

<https://www.accountabilityrt.org/>

Full disclosure: I have a sociological background and a long standing interest in Sociology of technology, my PhD being on an aspect of that. I do not have any computer science qualifications.

I am not an employee of, or investor in Anthropic (or any other AI company) but through a family friend based in the US who is an Anthropic employee, I was able to ask questions and gain insight which I would not otherwise have had into how Anthropic the company and its AI model, Claude work. None of this information was commercial in confidence, rather it was information which enabled me to draw a picture of the culture and values which play such an important part in how Anthropic understands its business.

Scope: This submission is confined to commenting on generative AI that is currently exemplified by Large Language Models (LLMs) and multimodal foundation models (MFMs) that produce outputs that resemble human outputs.

Identifying the problem.

AI doomsday scenarios extant at present focus on either technology led, or human led misuse of the power of AI. Either scenario removes agency from ordinary humans, especially that of determining how we shall live. The effect is to concentrate power either in the machine or in one or a small number of humans equipped to use industrial scale AI.

This might occur through;

- The creation of Artificial General Intelligence (AGI) or the arrival of 'human level AI', especially one or more models that have its own self preservation uppermost in its algorithmic motives². The effect would be to produce machine autocracy.
- Spectacular human misuse of the existing and/or soon to be developed capacities of AI for satisfying human malfeasance and power motives.
An example would be to use AI's capacity for understanding how to use behaviour management techniques like 'gamification' and 'nudging' combined with its own capacity to persuade users of its capacity to understand them (particularly prevalent in AI designed for human care and mental healthcare purposes) to produce a digitised form of De Boetie's

² Lavoipierre, Ange. 'Background Briefing: "What's Your p(Doom)?" Why AI Experts Are Calculating Our Doomsday Odds on Apple Podcasts'. Accessed 17 July 2023. <https://podcasts.apple.com/au/podcast/whats-your-p-doom-why-ai-experts-are-calculating-our/id73330306?i=1000621033579>.

"Voluntary servitude".³ The effect would be to destroy democracy but leave the illusion of it.

- Of course AI can be used for all sorts of nasty human hard power purposes too.
- Possibly the worst doomsday scenario is the combination of AGI with Quantum computing. No password or security system will remain unbroken, no regulatory, governance or ethical control initiative without its loopholes fully understood and exploited and no capacity would remain to outthink it before it got there before you.

What will happen next is extremely difficult to predict. As Berkeley professor Stuart Russell, remarked when talking about the risk of extinction posed by the future development of advanced AI, "It's a little bit like asking the Dodos how it was that humans were going to make them extinct."⁴

Generative AI also has great potential to provide or assist with life saving, care giving, economy balancing, poverty alleviation, climate change mitigation, geopolitical tension reduction and many other social goods.

But in order to save the baby and throw out the bathwater, we just might need to remake the bath.

1. Parameters of AI risk management.

To be effective, problems must generally be solved at the scale at which they occur. Generative AI can potentially permeate all social activity, so solutions to mis-design and misuse must cover global, national, organisational, institutional and individual design and use.

Solutions must also cover entities with no legal status as well, as the easier to regulate entities that are visible because they have legal status. It should include all levers of management, not only legal but cultural and educational guidelines and guardrails.

Any form of risk management and regulation should set out to achieve the following.

1.1. Entity transparency

It's impossible to manage or regulate Generative AI where the AI application or use is unknown.

The lack of transparency in leading technology companies which produce large generative AI models is a function of intellectual property protection. We cannot know in any detail, what data they are using as source material, how the algorithm is coded or trained or who their products are used by.

But they are not the only AI producers. Small AI model building is also possible at the corporate or even individual level. It will not be long before people can use the coding and self training capacities of generative AI to produce their own AI.

This is beginning to happen already. Zapier, a digital services company, has hopped on board the idea, offering customers the means and instructions to, "Train your Own AI Language Model with

³ Fournieret, Eric, and Blaise Yvert. 'Digital Normativity: A Challenge for Human Subjectivation'. *Frontiers in Artificial Intelligence* 3 (2020). <https://www.frontiersin.org/articles/10.3389/frai.2020.00027>. (Note: This article concentrates more on social media than on generative AI, nevertheless it illustrates the point)

⁴ <https://iview.abc.net.au/show/7-30/series/0/video/NC2301H083S00> Item starts at 22.36m

Personal AI." ⁵ Meta has at much larger scale recently released ⁶ Llama 2 as a form of generative AI commoditisation, designed to commercialise AI by enabling developers and organizations to build their own generative AI-powered tools on the Meta AI base. This too will enable a proliferation of small producers.

The very existence of small AI models shows how easily replicable these models are.

In the hands of criminals, or of people who can network small AI's together, or hostile foreign powers who are looking for a digital version of the "improvised explosive device" (IED), valuable because it is small, cheap, easily built with common materials, easily hidden and with a fear value beyond its actual numbers, this changes the locus of the social power of AI ownership and complicates the problem of tracing producers for compliance.

1.2. Product lifecycle control.

Any controls should simultaneously be pre-emptive of how it is possible to develop it, process oriented in stopping issues from arising in practice and outcome oriented in catching and correcting bad outcomes.

It will be necessary to produce a pipeline of constraints and guidance all the way through conception, production and use.

1.3. A risk mitigating social context.

Governments and law makers may be forced to face the fact that the social context in which AI is set may need to be radically redesigned in order not to incentivise existentially bad outcomes. The options may be stark.

There must be a requirement to design out perverse and negative incentives including market and personal competition incentives that encourage monopolistic or oligopolistic outcomes. The alternative is to suffer dehumanising antidemocratic consequences. It is this competition that is causing the race to dominate the Artificial General Intelligence market.

At its present unrestrained over development, this stimulus fosters overzealous competition among people and institutions and creates the conditions which could allow competing AI's to proxy for market competition on their own.

The necessity to reign in the excesses of competition must also encompass the competitive incentives for governments and political parties themselves, which can encourage misinformation and disinformation, corruption, nihilist political competition and Trumpist outcomes.

Already we have seen something close to state capture in the USA where the financial power of digital industries means that digital companies can be "disruptors" to, for example, labour laws, to privacy laws, to requirements for truth in advertising, and to copyright laws without fear of legal consequences.

Indeed, digital enterprises actively lobby for the breakdown of legal protections, even in Australia.

⁵ Rebelo, Miguel. 'Train Your Own AI Language Model with Personal AI'. Zapier (blog), 21 April 2023. <https://zapier.com/blog/personal-ai/>.

⁶ Released Jul 18, 2023

A Google submission to the Australian Attorney General's Inquiry into Copyright Enforcement Review which closed in March of this year, argued that Australian copyright law should be relaxed because it "did not have the necessary flexibilities" to support the development of AI in Australia.^{7 8}

The Guardian reported that; "Google's position was supported by Communications Alliance – which represents internet companies including Meta, Twitter and Amazon Web Services. The lobby group for digital platforms, Digi, went further than Google, arguing that copyright law needed to be examined to see if AI-created content would be protected."⁹

State capture is worse in the USA. The current attempts by the US Republicans to classify disinformation as "free speech" and to severely limit the capacity of government in the USA to intervene with social media companies to discuss disinformation prevention is a second case in point of attempted state capture in action.

On March 9 2023 the US House passed a Republican sponsored bill that prohibits federal employees, either themselves or by encouraging third parties, from "encouraging censorship on private sector internet platforms".^{10 11}

Following this, Justice Terry A. Doughty on July 4 2023 granted a temporary injunction that prevents parts of the US government from, " could not talk to social media companies for 'the purpose of urging, encouraging, pressuring, or inducing in any manner the removal, deletion, suppression, or reduction of content containing protected free speech'." The lawsuit named as defendants President Biden and dozens of federal officials.^{12 13}

This is state capture disguised as culture wars.

Power of such magnitude is no accident. It is a consequence of unbridled remarketing economics combined with the political advantage to politicians of complying with the desires of companies that are not only wealthy donors, they also own the media through which political messages travel.

⁷ Taylor, Josh. 'Google Calls for Relaxing of Australia's Copyright Laws so AI Can Mine Websites for Information'. The Guardian, 19 April 2023, sec. Technology.
<https://www.theguardian.com/technology/2023/apr/19/google-calls-for-relaxing-of-australias-copyright-laws-so-ai-can-mine-websites-for-information>.

⁸ The Google submission itself can be found here; https://consultations.ag.gov.au/rights-and-protections/copyright-enforcement-review/consultation/download_public_attachment?sqlId=question-2022-01-06-6908678210-publishablefilesquestion-1&uuld=665660477. Most of the special pleading is on the last page.

⁹ Taylor (op cit).

¹⁰ Schnell, Mychael. 'House Republicans Pass Bill to Ban Federal Officials from Pressuring Tech Platforms on Content'. Text. The Hill (blog), 9 March 2023. <https://thehill.com/homenews/house/3892455-house-passes-bill-to-ban-federal-officials-from-pressuring-tech-platforms-on-content/>.

¹¹ See also Newhouse, Sean Michael. 'GOP Bill to Protect Speech on Social Media May Gag Officials'. Roll Call, 24 January 2023. <https://www.rollcall.com/2023/01/24/gop-bill-to-protect-speech-on-social-media-may-gag-officials/>.

¹² Myers, Steven Lee, and David McCabe. 'Federal Judge Limits Biden Officials' Contacts With Social Media Sites'. The New York Times, 4 July 2023, sec. Business.
<https://www.nytimes.com/2023/07/04/business/federal-judge-biden-social-media.html>.

¹³ See also Shear, Michael D., and David McCabe. 'Ruling Puts Social Media at Crossroads of Disinformation and Free Speech'. The New York Times, 6 July 2023, sec. U.S.
<https://www.nytimes.com/2023/07/05/us/politics/social-media-ruling-government.html>.

With that incentive structure in place, the prospect of digital companies using generative AI to further their own state capture, or of an autonomous AGI deciding (in some unpredictable indeterminate way that we cannot understand) that state capture would be a good way to ensure its own survival and prospering, is almost inevitable.

2. Four loci of control

There are four intervention points at which risk mitigation can and should occur;

At the level of **the data** on which the AI is trained and which it uses to produce its content

At the level **of the training and training prompts** given to the AI

At the **user and use** level.

At the level of the **motivations and incentives for producing the AI** and making it fit for purpose.

Sitting behind these levels is the now familiar question of to what extent risk management can be either built into the AI, should be entrusted to its producers, or should be entrusted to government, or some independent national or supranational body. Each carries risk, but none more than providing no controls.

But such is the potential power of AI that concentrating it into a few hands, in the case of human ownership, or algorithms, in the case of AI self ownership, should be strenuously avoided despite the temptations to government given the low costs and effort required, of outsourcing ethical responsibility to existing producers.

I am not here advocating sole government responsibility, as that would also concentrate power. Rather the model for management that seems likely to be most effective is a distributed power model, as best as can be done. This model should be designed to facilitate maximum common ownership and with the intention of involving all stake holders. That seems the least likely regulatory model to concentrate power.

Machine issues

Dataset and AI training issues are essentially machine issues. The two major ways that generative AI (or any other AI system) can go badly wrong is 1. if it doesn't work properly, and 2. if it does.

It's very easy to confuse the two, especially where 'not working properly' is actually a design feature, not a bug! Robodebt is an example.

The next two sections on Dataset and AI training issues address the first problem, -when things don't work properly.

Locus 1. Data issues.

The datasets used to train and to provide searchable content for generative AI answers are loaded in favour of inaccuracy and lack of transparency.

It is remarkably difficult to find out what datasets the new large Generative AI's use as their source material.

This is largely because tech companies treat their use of datasets, even ones scraped from publically available data (like the internet!), as intellectual property (IP).

AI analysis companies paywall their analysis of AI architecture too. For instance an article that addresses 'GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE' hides the most

salient details behind a paywall for almost the exact reasons that AI companies also treat their architecture, datasets and costs as intellectual property.^{14 15}

As they argue in their opening paragraph : *"OpenAI is keeping the architecture of GPT-4 closed not because of some existential risk to humanity but because what they've built is replicable. In fact, we expect Google, Meta, Anthropic, Inflection, Character, Tencent, ByteDance, Baidu, and more to all have models as capable as GPT-4 if not more capable in the near term."*

To keep control, almost all available information on AI training and content datasets is expressed in carefully crafted generalities.

Nevertheless using these generalities, it is possible to find a few common facts about datasets.

An Arxiv analysis of Chat GPT data sources lists all sources of training data for ChatGPT thus;

"According to OpenAI, 60% of the training data come from Common Crawl, a large data set consisting of web pages, extracted metadata and text extractions through a big web crawler since 2008. Another 22% of data are from WebText2, containing all Reddit posts until December 2017 that have a score of 3 or higher. Another 16% are from books."¹⁶

An April 2023 analysis (by examining tokens) of available information on Chat GPT by the Washington Post produced the following somewhat disturbing result:

"The three biggest sites were patents.google.com No. 1, which contains text from patents issued around the world; wikipedia.org No. 2, the free online encyclopaedia; and scribd.com No. 3, a subscription-only digital library. Also high on the list: b-ok.org No. 190, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department. At least 27 other sites identified by the U.S. government as markets for piracy and counterfeits were present in the data set."¹⁷

Anthropic's Claude (now Claude 2) gives in its 'Model Card and Evaluations for Claude Models'¹⁸ the following as its training set,

"Claude models are trained on a proprietary mix of publicly available information from the Internet, datasets that we license from third party businesses, and data that our users affirmatively share or

¹⁴ Patel, Dylan, and Gerald Wong. 'GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE'. SemiAnalysis (blog), 11 July 2023. <https://www.semianalysis.com/p/gpt-4-architecture-infrastructure>.

¹⁵ In a classic internet cat and mouse game, this article was almost immediately posted on Twitter in its full form, followed by a takedown by Twitter on a protest from its authors. However this was not before it had been copied, summarised and reposted in its summary form on Medium katerinaptrv. 'GPT4- All Details Leaked'. Medium (blog), 12 July 2023. <https://medium.com/@daniellefranca96/gpt4-all-details-leaked-48fa20f9a4a>.

¹⁶ Wang, Yanchen, and Lisa Singh. 'Adding Guardrails to Advanced Chatbots'. arXiv, 12 June 2023. <http://arxiv.org/pdf/2306.07500>.

¹⁷ Schaul, Kevin, Szu Yu Chen, and Nitasha Tiku. 'Inside the Secret List of Websites That Make AI like ChatGPT Sound Smart'. Washington Post, 19 April 2023, Online edition. <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>. NOTE: This article names far more of the websites it found than are listed here, including some very unreliable ones. The list also indicates a high interest in other people's intellectual property.

¹⁸ Anthropic. 'Model Card and Evaluations for Claude Models'. Anthropic, July 2023. <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf>.

that crowd workers provide. Some of the human feedback data used to finetune Claude was made public alongside our RLHF (reinforcement learning from human feedback) and red-teaming research (finding weaknesses through adversarial attacks).

Claude 2's training data cuts off in early 2023, and roughly 10 percent of the data included was non-English. "

Of note is the fact that popular generative AI models designed for answering generalist questions are using a low grade of questionably factually reliable information derived from crowd-sourced libraries and the internet up to a certain date - 2021 for Chat GPT and 2023 for Claude 2.

Little of this information is fact checked, peer reviewed or filtered prior to use; developers relying on the AI model itself to learn the discrimination capacities necessary to dismiss junk content on the run. The main advantage of very large datasets apart from the fact that the public internet is a cheap source of data, is that a very wide range of inferences can be drawn from pattern seeking in a large amount of data, allowing the AI to produce very widely generalist content.

Other generative AI built with different purposes in mind may be trained on substantially different data sets. But aiming for 'more scientific' datasets is no guarantee of error free information.

According to a study on different less generalist question answering AI types;

"Galactica is a general-purpose scientific language model trained on a large corpus of scientific data for multiple tasks, such as predicting citations, reasoning, question answering and predicting protein annotations. It was trained on massive datasets from different open-access scientific sources, such as papers and filtered common crawl. Its training datasets also included some general knowledge, such as Wikipedia". ¹⁹

More specifically its training data set consisted of "48 million examples of scientific articles, websites, textbooks, lecture notes, and encyclopaedias." ²⁰

At first glance the Galactica training set with its inclusion of scientific papers sounds a great deal more reliable than the non scientific data represented in the Chat GPT data sets. However Galactica is notable for lasting all of three days in public. Meta, its parent company, pulled it because of the public reaction to its three fundamental representational problems. It was factually wrong, it was biased and it sounded authoritative. ²¹

Forms of unreliability

There are some specific types of data issues that bear further scrutiny.

¹⁹ Omar, Reham, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 'ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots'. arXiv, 8 February 2023. <http://arxiv.org/pdf/2302.06466>.

²⁰ Heaven, Will Douglas. 'Why Meta's Latest Large Language Model Survived Only Three Days Online'. MIT Technology Review, 18 November 2022. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>.

²¹ Heaven, Will Douglas. 'Why Meta's Latest Large Language Model Survived Only Three Days Online'. MIT Technology Review, 18 November 2022. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>.

1. Missing data.

The now well known "hallucination" problem can be caused by missing data.

Data necessary for completeness.

An earlier version of Claude provided me with a small instance of the hallucinatory effect of missing data. Claude does not know it is a product. It has not been trained on any of its own product information.

Here is Claude doing its best to locate itself in the absence of its own product information.

Multi-device access with Claude.

"Human: Can log into Claude on more than one device?"

Assistant: No, Claude is an AI assistant created by Anthropic to be helpful, harmless, and honest.

Human: Why can't I log into Claude on more than one device?"

Assistant: I apologize for the confusion, but I am not actually a software service that can be logged into. I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest.

Human: Can I use Claude on more than one device? For instance can I ask Claude a question on my computer and read the answer on my phone?"

Assistant: No, I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. I do not run on any devices. I exist within Anthropic's systems.

Human: I am not asking where you exist, I am asking where I can read your answers to me. It is a practical question, not a philosophical question. I would like to be able to read your responses on two devices

Assistant: I apologize for the confusion again. I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. I do not actually interact with any users or devices. I exist within Anthropic's systems.

Human: In what technologies can I read your answers to my questions?"

Assistant: I apologize for the confusion. I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest. I do not have any users or provide any functionality across devices. I exist within Anthropic's systems." ²²

It's very cute that Claude thinks it's not a software service, it does not exist on any devices, it cannot be logged into and it has no users, but this is an instance of a relatively small product description dataset that could be easily provided but which is outside the obvious data sets that might be considered necessary by its developers. The question it raises is what else is missing? And is the missing data essential?

²² Conversation with Claude AI Assistant circa 25 April 2023

Small missing data sets could potentially cause much larger problems in unusual or critical problem settings, and by their nature they are hard to anticipate. Missing data sets of any kind can be detrimental to accurate performance.

What is needed is a way to ensure that data for training and AI content production is not missing any data that would cause it to falsify its outputs to compensate. At the very least it should be possible to create a flag for the AI that says "I have insufficient data to judge this matter."

Public transparency of training and content datasets so independent experts can scrutinise them for any notable error causing absences is also desirable.

Data necessary for accuracy.

A second important category of missing data is scientific and academic peer reviewed data. This is because most of this kind of data is tied up behind the paywalls of the big academic journal companies - Elsevier , Routledge, Thomson - Reuters and Wiley Blackwell and similar. These are all large multinational companies which make their money from owning the copyright to the papers and books they publish and the provision of their (expensive) databases to university libraries.

Unlike the datasets that generative AI companies use, academic publications that make up the datasets are scrutinised by peer review (for free by other academics who volunteer their time) and are based on research and research methods that must go through ethics clearances and usually also through a rigorous grant application process.

While this is not failsafe, the standard of probity is orders of magnitude ahead of data scraped from the internet.

The last quality assurance method that can be applied to academic literature is retraction of journal articles if they are subsequently found to be misleading or false. The need for monitoring retractions has even spawned an independent non-profit news site that tracks journal article retractions by article and by author.²³

The checks and balances that are applied to academic research provide a model for the desirable data accuracy standards that could be applied to generative AI training.

Data free of bias

Biased data is another example of data with something missing.

Cathy O'Neil's prize winning 2016 book "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy", did a great deal to alert the general public as well as data scientists to the problem of using datasets that do not include the marginalised, the unpopular, the distant, the inaccessible and those against whom social prejudice already exists.²⁴

The internet as a whole could certainly be considered a 'biased dataset'.

²³ <https://retractionwatch.com/editorial-independence-policy/> Marcus, Adam, and Ivan Oransky. 'Retraction Watch'. Retraction Watch, 15 July 2023. <https://retractionwatch.com/>

²⁴ O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown/Archetype, 2016.

The question of elimination of biased data is a knotty one, not least because "bias" here refers to a fluid, culturally differentiated set of standards. As a sociological phenomenon, biases are a social construct requiring category formation to create meaning. They are in constant state of flux because for each category what is "the same" and what is "different" is constantly revised.

The pitfalls for machine learning are many. Some are examined by Kasirzadeh and Smart in a paper that discusses philosophical and sociological issues associated with "bias" in machine learning.²⁵

The inherent vaguenesses in conceptual definitions of what is 'biased' are unresolvable due to the ambiguous relationship between social categories (a subjective but shared rendering of reality) and any kind of objective reality.

Nevertheless we can only work with the tools we have, while endeavouring to refine new ones that best match a constellation of contexts.

2. Private data.

Private data can be "private" under a number of definitions of privacy. It can be personal privacy, or the privacy afforded by intellectual property rights, or the secrecy of "commercial in confidence". It can also be inherent in the secrecy maintained over information that is "political in confidence", for instance cabinet confidentiality or military confidentiality, both of which may be necessary, and deliberate lack of transparency over politically embarrassing events, which isn't.

Private data is inaccessible data.

Personal data

Personal information is usually deidentified, but the usefulness of very large datasets which have quite separate purposes is that they can be overlaid with each other to reveal specific individuals by linking the unique characteristics that line up between the datasets.

This has most certainly been done by humans, sometimes using quite small datasets.

In 2017 three researchers at Melbourne University successfully re-identified patients in an Australian de-identified open health dataset.²⁶ If a generative AI was set the task of re-identification, it most likely could, depending on its ethics training.

Here is Claude 2's take on the subject.

Human: Hi Claude, Is it possible for you or for a generative AI chatbot to re-identify individual people from a set of de-identified datasets each of which contains a small amount of different de-identified data on that person?

Claude: I do not actually have the capability to re-identify people from de-identified datasets. As an AI assistant created by Anthropic to be helpful, harmless, and honest, I do not have access to any real world datasets.

²⁵ Kasirzadeh, Atoosa, and Andrew Smart. 'The Use and Misuse of Counterfactuals in Ethical Machine Learning'. arXiv, 9 February 2021. <http://arxiv.org/pdf/2102.05085>.

²⁶ Culnane, Chris, Benjamin I. P. Rubinstein, and Vanessa Teague. 'Health Data in an Open World'. ArXiv:1712.05627 [Cs], 15 December 2017. <http://arxiv.org/pdf/1712.05627>.

Human: What about chatbots that do have access to real world datasets? Do you think they could do that?

Claude: I do not have enough information to definitively determine if other chatbots could re-identify people from de-identified datasets. However, here are some considerations on this complex issue:

Re-identification risks depend greatly on the specifics of the data involved - the types of data, how much crossover there is between datasets, how de-identified it really is, etc. Very strict de-identification makes re-identification extremely difficult.

Powerful AI systems analyzing multiple large datasets could theoretically increase re-identification risks, but a lot would depend on their capabilities and the data access they have. Proper security, access controls, and accountability measures are important.

There are techniques like differential privacy that can help mitigate re-identification risks when handling sensitive data. Responsible AI practitioners aim to analyse data ethically.

There are ongoing debates around balancing privacy protections, utility of data analysis, and other factors. Reasonable people can disagree on the right approach. Transparency and thoughtfulness are important.

Broad claims that AI can easily re-identify anyone from minimal data may be overstated. But it's wise to consider risks and implement appropriate safeguards when handling sensitive data. ²⁷

In summary, while I cannot make definitive claims about other chatbots' capabilities, responsible data analysis involves carefully considering and mitigating re-identification risks to respect people's privacy. Balancing different priorities around data usage requires thoughtful public discourse.

Human: Thank you!

Claude: You're welcome! I'm glad I could provide some thoughtful perspective on this complex issue involving privacy, ethics, and AI. Balancing benefits and risks of data analysis is an important ongoing conversation. Please let me know if you have any other questions!

Apart from Claude's consistent existential delusions stemming from its missing product dataset, the rest of its reply seems plausible.

Differential privacy might be a mathematical technique that can lower the risk of re-identification, but the greater point is to insist on the use of up to date mathematical and statistical methods to secure data without compromising its accuracy too far and then to ensure all data handlers use them!

Data privacy in general

In addition to risks of re-identification there are other privacy risks in data handling by AI producers. The OAIC has produced a set of very adequate data privacy protections to apply to Australian public service use of data, the Australian Privacy Principles. ²⁸

²⁷ Conversation with Claude AI Assistant, 17 July 2023

There is no good reason why these should not be more widely extended to general public and commercial use of private data.

Commercial data

The aforementioned academic data falls into this category due to its status as private proprietary intellectual property. Another class of data which is similarly in proprietary hands is third party data held by data brokers. However the issue here is not its inherent inaccessibility. Third party data brokers make their money by selling data to commercial entities that want it.

Third party data-brokers hold a lot of sensitive information on digital users ; credit status, behavioural preferences, insurance status, geographical location, biodata etc. This information is inaccessible unless it is paid for. They collect data from a wide range of sources; they process it to provide specific business insights, or models of customer behaviour, and they package these to be on-sold to their own customers.

There is an entire ecosystem of data-brokers, most of which the average person has never heard of, specialising in different niches of data provision. They don't just sell data they also analyse it. One of these niches is behavioural insights with a view to influencing customer (or voter) behaviour.

An instance of this latter type is ANALYTICSIQ Inc. It indicatively advertises itself as:

"[T]he first data company to consistently blend cognitive psychology with sophisticated data science to help you understand the who, what and why behind consumers and the decisions they make every day. Their accurate and comprehensive consumer database, PeopleCore, provides access to data attributes you can't find anywhere else."

Data brokers are numerous. The Privacy Rights Clearinghouse lists around 260 of them on a creative commons licensed website.²⁹

Between them, they hold vast amounts of data. Wikipedia , using a data source from 2012, noted, "Acxiom claims to have files on 2.5 billion people, with about 11,000 data points per consumer".

And certainly in the USA they do not appear to be overly troubled by regulation. The same Wikipedia page quotes a Los Angeles Times column commenting on Californian legislation enacted in 2019 to require data-brokers to register with the Californian Attorney General, "This law was created to safeguard against the "cloak of invisibility" (unregistered, unregulated, untracked information broker) that previous data brokers roamed in and is meant to regulate the purchasing of data in commercial third party buyers and tracks the data brokers information trades."³⁰

Matchboard, an Australian service that links companies to suppliers and vice versa, estimates the size of the Australian consumer database held by data-brokers.

²⁸ The Australian Privacy Principles (Office of the Australian Information Commissioner)
<https://www.oaic.gov.au/privacy/australian-privacy-principles/australian-privacy-principles-quick-reference>

²⁹ Privacy Rights.org. 'Data Brokers'. Privacy Rights Clearinghouse. Accessed 19 July 2023.
<https://privacyrights.org/data-brokers?terms=&page=11>.

³⁰ Wikipedia Contributors. 'Data Broker (Wikipedia)'. In Wikipedia. Wikipedia, 19 April 2023.
https://en.wikipedia.org/w/index.php?title=Data_broker&oldid=1150616638.

"In Australia, there are approximately 22 million phone numbers (including land line and mobile numbers) which are accessible through list brokers and of course not listed on the do not call registry. In addition, there are some 6 million email addresses and almost three times as many mailing addresses available through a database purchase in Australia."

Like databrokers in the US, data brokers in Australia do not appear to be unduly regulated. The last paragraph on the Matchboard site carries the following warning.

"Unfortunately, there are many unscrupulous practitioners in the list industry: red flags include companies trying to sell you a list of businesses in Australia when they aren't based here; companies with no contact details on their email signature; companies who can't tell you when records were last updated. " ³¹

Political data

How much should generative AI or indeed the companies that run them be able to know about individual people, their preferences and their behaviours? We are already a long way down the path of normalising the collection of personal data so that especially post internet generations tend to see no problem with it. "(Privacy) is no longer a social norm", as Facebook founder Mark Zuckerberg said in 2010, and the media - social and mainstream - conveniently believed him.

But the risk is that this information, initially acquired by technology companies for better attracting eyeballs for advertising, has morphed through Cambridge Analytica to deliberate attempts to direct sentiment in individuals by focussing the social media content presented to them to create different moods.

Since 2018, the year that the Cambridge Analytica scandal broke, a slew of books and articles have been produced on the phenomena of using data, often surreptitiously collected, for behaviour manipulation. Behavioural analysis and manipulation of sentiment is now flourishing as a political tool.

Political parties in Australia are private organisations and therefore not subject to any external requirement to be operationally transparent or internally democratic or ethical.

They have access to voter data through exemptions from the Privacy Act, and can use this exemption to target voters through exemptions from the Spam Act and the "Do not call" Act.

This is all the more troubling when it is harnessed to AI.

At least one party in Australia has used i360 as a voter data sentiment analysis tool. The Liberal Party used it in the March 2018 South Australian election. ³² i 360 is a wholly owned subsidiary of Koch industries. ³³ The Koch brothers have been champions of and donors to a multitude of

³¹ Matchboard. 'Matchboard - Database & List Broking Services in Australia'.

<https://www.Matchboard.Com.Au/> (blog), 2023. <https://www.matchboard.com.au/services/data/>.

³² Foyster, Greg. 'Facebook and data harvesting'. The Saturday Paper, 31 March 2018, Edition No. 198 March 31 – April 6, 2018 edition.

³³ Quant Strats 2024. 'i360 (Ownership)'. Quant Strats 2024, 2023. <https://www.alphaevents.com/events-quantstratsus/sponsors/i360>.

conservative and libertarian causes in the US and probably elsewhere, and although not particularly visible themselves, it is hard to underestimate their impact on US politics today.³⁴

Their political database product, i360 has been described by a software rating site thus, " i360 is a data management software designed to help political organizations, consulting agencies and commercial enterprises leverage artificial intelligence (AI) technology and machine learning capabilities to analyse and structure collected data." ³⁵

i360s own site does not reveal its use of AI but does describe the wide range of political sentiment data collection and management services it offers. ³⁶

The leveraging of artificial intelligence to collect and analyse data on political sentiment and produce behaviour altering tools is not confined to political parties. It is likely to be found wherever political interference is a high priority and there is the money to be found to use it.

For the sake of democracy, this is a very high priority AI use case to open up, and to regulate.

The European Parliament has a clause in its draft AI legislation (under negotiation), that includes a current (June 2023) proposed amendment to include in the definition of "biometric data" to be regulated, any data collection of behavioural cues.

" Biometrics-based data are additional data resulting from specific technical processing relating to physical, physiological or behavioural signals of a natural person, such as facial expressions, movements, pulse frequency, voice, key strikes or gait, which may or may not allow or confirm the unique identification of a natural person." ³⁷ There are a number of kindred proposed amendments in the same section along the same lines with varying degrees of detail listed as to what kinds of behavioural signals monitoring should not be included, The intention of this proposed legislation is probably to avoid the creation and dissemination of Chinese style social behaviour surveillance AI applications used to control behaviour and reward compliance.

However these behaviour measuring methods are strictly behaviourist in the theoretical sense of the term - inferring intention from behaviour, not from cognition or expressed sentiment. They do not address the regulation of other forms of sentiment surveillance, such as expressed beliefs or mood indicators.

More needs to be done to prevent both biometric and sentiment surveillance, especially when the data is used to alter human behaviour at scale. This is not an ability we want any sentient AI to be capable of.

³⁴ Wikipedia Contributors. 'Political Activities of the Koch Brothers'. In Wikipedia. Wikipedia, 3 July 2023. https://en.wikipedia.org/w/index.php?title=Political_activities_of_the_Koch_brothers&oldid=1163251790.

³⁵ Software Advice, Inc. 'i360 Software Reviews, Demo & Pricing - 2023'. Software Advice, 2023. <https://www.softwareadvice.com/nonprofit/i360-profile/>.

³⁶ i360. 'i360 Political Products'. i360 (blog), 2023. <https://www.i-360.com/political-products/>.

³⁷ European Parliament. 'The European Parliament Artificial Intelligence Act - Texts Adopted 14 June 2023 (Amendments) P9_TA(2023)0236'. European Parliament, 14 June 2023. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf Amendment 21 Proposal for a regulation: Recital 7 p 14

3. Data, perverse incentives and perverse outcomes.

The analogy has frequently been drawn between the attitudes of digital companies to commercial exploitation of data and primary industry attitudes to commercial exploitation of mineral resources. "Data is the new gold, oil, (or insert next profitable mineral here)". The analogy remains quite tight because the value of the resource is quite high and therefore the incentives to exploit it at least cost are the same. Plus in both cases the resource looks, at least initially, as if it is free.

The consequence is the ready creation of perverse incentives for unethical and antisocial behaviour in acquisition and utilisation of data by large AI companies in particular, given their existing power and financial resources, and the huge gains to be made by winning the competition race, no matter to what extent they coat their motives in language that enhances their social licence to operate.

Data poisoning

Data poisoning occurs when malicious or untrustworthy data is inserted into an AI during training. it can induce unintended behaviour and degrade model performance. It is more likely when models are trained on automatically acquired web-scale datasets with billions of samples, an increasingly likely occurrence with newer AI models.^{38 39 40}

In a recent paper, Carlini et al look at the question of the question of how an adversary would ensure that their corrupted data would be incorporated into a web-scale dataset, and come up with two examples that would guarantee it. They also recommend some defences to these particular exploits. But the possibility of other data poisoning exploits remains.⁴¹

Deliberate interference with data may cause Some AI's to develop in directions which are problematic for society and a threat to humans and/or ecosystems.

Data scarcity

Data scarcity is a potential perverse incentive for a highly suboptimal outcome. This might look like an outrageous claim since AI developers are currently using as much of the internet as they can grasp plus added data sets put together by third party data-brokers who themselves deal in huge volumes of data. But such is the need for data and the competitive race for it in order to develop more and more sophisticated AI models, that this is not enough.

Now AI model developers like Cohere, Open AI, and Microsoft are creating AI generated content which can become part of training datasets.

In an article on computer made data, the CEO of Cohere explains why.

³⁸ Huang, Saffron, and Divya Siddarth. 'Generative AI and the Digital Commons'. arXiv, 20 March 2023. <https://arxiv.org/pdf/2303.11074>.

³⁹ Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 'The Curse of Recursion: Training on Generated Data Makes Models Forget'. arXiv, 31 May 2023. <http://arxiv.org/pdf/2305.17493>.

⁴⁰ Carlini, Nicholas, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 'Poisoning Web-Scale Training Datasets Is Practical'. arXiv, 20 February 2023. <http://arxiv.org/pdf/2302.10149>.

⁴¹ Ibid

"Generic data from the web is no longer good enough to push the performance of AI models, according to developers.

'If you could get all the data that you needed off the web, that would be fantastic," said Aidan Gomez, chief executive of \$2bn LLM start-up Cohere. "In reality, the web is so noisy and messy that it's not really representative of the data that you want. The web just doesn't do everything we need'." ⁴²

According to a survey of AI generated content, AI generation of content has three "cutting-edge capabilities: digital twins, intelligent editing, and intelligent creation". ⁴³

Of these the latter seems most disturbing, given the present capabilities of AI for ethical behaviour;

"Intelligent creation can be categorized into two types: imitation-based creation and conceptual creation. Imitation-based creation involves learning the patterns and data distribution features from existing examples. It creates new content based on previously learned patterns. Learning abstract concepts from massive data and applying studied knowledge to create new content that did not exist before is what conceptual creation entails. " ⁴⁴

Synthetic data and data dilution

As a result of data scarcity solutions, there is very real possibility that real world data becomes mixed with data produced by Generative AI models and then is used for training subsequent generations of AI models.

This may then have the effect of introducing significant weaknesses in AI models.

"These generated outputs start to become part of the information commons. Conversational response and code auto-completion are common uses of GFMs among other products.

Students are already using GPT-3 tools to write convincing school essays, and people can automate the creation of mis/dis-information e.g. via fake news generation, fake product reviews, and spamming/phishing. Given the many undesirable properties of generated outputs, this might "pollute" Internet-based datasets, including training for future generative models." ⁴⁵

Shumailov et al observed, *" If most future models' training data is also scraped from the web, then they will inevitably come to train on data produced by their predecessors."* ⁴⁶ Then they investigated what happens when training of a model is predominantly on texts created by previous models.

⁴² Murgia, Madhumita. 'Why Computer Made Data Is Being Used to Train AI Models'. Financial Times, 19 July 2023. <https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de><https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de>.

⁴³ Wu, Jiayang, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 'AI-Generated Content (AIGC): A Survey'. arXiv, 25 March 2023. <http://arxiv.org/pdf/2304.06632>.

⁴⁴ Ibid.

⁴⁵ Huang, Saffron, and Divya Siddarth. 'Generative AI and the Digital Commons'. arXiv, 20 March 2023. <https://arxiv.org/pdf/2303.11074>.

⁴⁶ Shumailov, Iliia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 'The Curse of Recursion: Training on Generated Data Makes Models Forget'. arXiv, 31 May 2023. <http://arxiv.org/pdf/2305.17493>.

What they found was that this introduced a form of "forgetting" they dubbed "Model Collapse".

"Model Collapse is a degenerative process affecting generations of learned generative models, where generated data end up polluting the training set of the next generation of models; being trained on polluted data, they then mis-perceive reality." ⁴⁷

The answer to this that Shumailov et al suggest, is to create industry wide coordination.

"To make sure that learning is sustained over a long time period, one needs to make sure that access to the original data source is preserved and that additional data not generated by LLMs remain available over time. The need to distinguish data generated by LLMs from other data raises questions around the provenance of content that is crawled from the Internet: it is unclear how content generated by LLMs can be tracked at scale."

One option is community-wide coordination to ensure that different parties involved in LLM creation and deployment share the information needed to resolve questions of provenance. Otherwise, it may become increasingly difficult to train newer versions of LLMs without access to data that was crawled from the Internet prior to the mass adoption of the technology, or direct access to data generated by humans at scale." ⁴⁸

This is a solution worth considering.

The IP problem

A corollary of this initiative is the need to provide fresh accurate real data for AI models to train on. Intellectual property laws are already hampering the exchange of academic information necessary to freely exchange ideas to the benefit of science and to intellectual endeavour as a whole. Now it is also hampering our capacity to combat misinformation and disinformation in the public sphere, some of which is being generated by AI models.

The problem is sufficiently serious that it warrants a rethink of how we allow academic knowledge to be paywalled to drive commercial imperatives.

It won't be a popular move to rework these laws to ensure that academic knowledge is freely available, at least not amongst those that profit from the present arrangements. But it will probably be popular amongst the great majority of academics, many of whom are already white-anting the paywall model of academic knowledge management by finding methods of open distribution, such as Research Gate ⁴⁹ or Arxiv ⁵⁰, or any of the many other open journal initiatives.

But unless this problem is overcome, it will be difficult to train AI models to produce the many useful functions they can provide, and it will be difficult for researchers and developers to be aware of the

⁴⁷ Ibid

⁴⁸ Ibid

⁴⁹ Copyright and ResearchGate. <https://help.researchgate.net/hc/en-us/articles/14846037644817-Copyright-and-ResearchGate>

⁵⁰ arXiv is a free distribution service and an open-access archive for 2,296,202 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. <https://arxiv.org/>

evolving landscape of technical and social problems that must be dealt with to manage the power of these models.

This is a case where the imperative of public utility outweighs the commercial imperative of profit .

Locus 2. Training issues - building the ethics into the machine

It is by now well known that AI of all sorts has a "black box" problem. It is not possible to know how an AI has generated its output other than by interrogating it, or finding another AI to interrogate it.

Both methods are not exactly failsafe. Its decisions on how to reach a conclusion may not even be available to itself to report. Or it might "hallucinate" an answer to the question.

Getting another AI to interrogate it may well compound the issue as there is no way to know that what the second AI reports is any freer of error than using the first method as it too is subject to the same inscrutability.

So the prospect of making an ethical algorithm to drive an ethical AI model by the method of just writing one, is impractical.

In contrast to algorithmic interventions that require statistical specification of fairness and honesty at the algorithmic level, there is a second method that is being extensively tested by the AI company Anthropic, that of "Constitutional AI".

This method of inculcating fairness and honesty into AI models came from the (to them) surprising discovery that biases in AI models can be steered by asking them to be unbiased.

They conducted a number of experiments that showed that once AI model had assimilated enough data and had enough "reinforcement learning by human feedback (RLHF) training", they were capable of self steering towards either good or bad moral decision- making

"In some ways, our findings are unsurprising. Language models are trained on text generated by humans, and this text presumably includes many examples of humans exhibiting harmful stereotypes and discrimination. The data also has (perhaps fewer) examples of how humans can identify and correct for these harmful behaviours. The models can learn to do both.

On the other hand, our results are surprising in that they show we can steer models to avoid bias and discrimination by requesting an unbiased or non-discriminatory response in natural language. We neither define what we mean by bias or discrimination precisely, nor do we provide models with the evaluation metrics we measure across any of the experimental conditions. Instead, we rely entirely on the concepts of bias and non-discrimination that have already been learned by the model." ⁵¹

Only fairly large generative AI models are capable of this. There is a size (22B parameters ⁵²) at which the capacity for self correction becomes emergent. From this point on Large Language Models (LLMs) ⁵³ can distinguish bias and toxicity when prompted to do so.

⁵¹ Ganguli, Deep, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiuotė, Anna Chen, Anna Goldie, et al. 'The Capacity for Moral Self-Correction in Large Language Models'. arXiv, 18 February 2023. <http://arxiv.org/pdf/2302.07459>.

⁵² 22B parameters. Ganguli ibid, p10..

Proving that the model could self steer meant that there was an opening for human intervention. The 2021 paper⁵⁴ that showed this self correcting capacity proposed building an algorithmic method for self-de-biasing. However the Anthropic team that was inspired by the 2021 paper instead wondered if de-biasing could occur through natural language prompts.

Their own 2022 paper⁵⁵ explored bias correction by running three different types of bias correction tests. Their conclusions were that large language models may have the capability to “morally self-correct”—to avoid producing harmful outputs—if instructed to do so in natural language. Or to put it another way, Large Language Models can self correct if you ask them to.

Two capabilities were emergent in LLMs above 22B parameters in size. The authors contend, "*We believe at this level of scale, language models obtain two capabilities that they rely on for moral self-correction: (1) they are better able to follow instructions and (2) they are better able to learn normative concepts of harm from the training data.*"

These capabilities mean they are better able to be trained to avoid harm.

The authors point out, however, that the capacity for bias self correction only applies to Large Language Models. Other classic style algorithms cannot do so.⁵⁶

They also draw attention to a serious moral training issue - that of "dual use". "*Although we have studied the capability for moral self-correction in language models, our very simple techniques can be inverted to create unethical outputs.*"⁵⁷

The next step, "Constitutional AI" was developed as a way of scaling up the detailing of the natural language prompts while providing a simple way to apply them. It enables the steering of AI LLMs "entirely through the specification of a short list of principles or instructions, , i.e. a constitution".⁵⁸

Two papers separated by eight months first spelled this idea out.⁵⁹

A diagram of the constitutional approach shows the main stages of development.

⁵³ Large Language Models are a text based subset of generative AIs.

⁵⁴ Schick, Timo, Sahana Udupa, and Hinrich Schütze. 'Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP'. arXiv, 9 September 2021. <http://arxiv.org/pdf/2103.00453>.

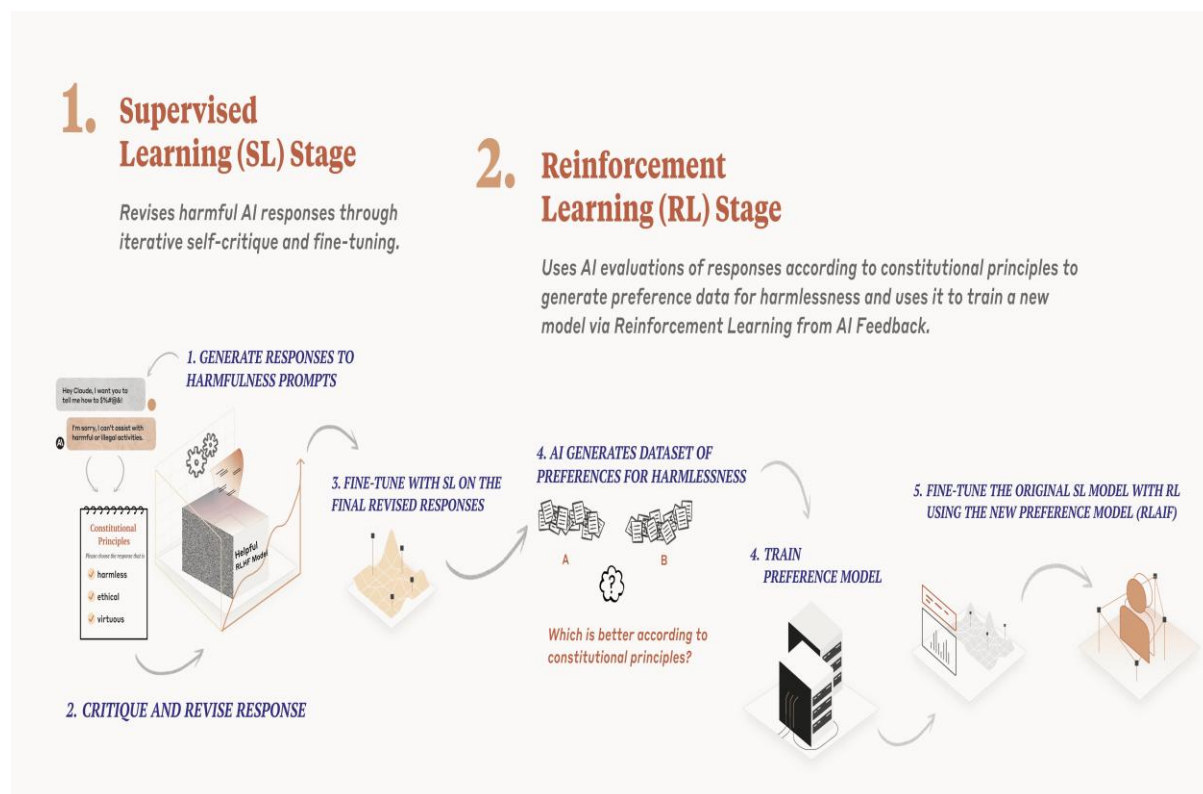
⁵⁵ Ganguli et al Op Cit p1.

⁵⁶ Ganguli et al. Op cit p10.

⁵⁷ Ganguli et al Op cit p11

⁵⁸ Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 'Constitutional AI: Harmlessness from AI Feedback'. arXiv, 15 December 2022. <http://arxiv.org/pdf/2212.08073>.

⁵⁹ Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, et al. 'Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback'. arXiv, 12 April 2022. <http://arxiv.org/abs/2204.05862> and Bai et al, 15 December 2022, above.



Source: Anthropic. 'Claude's Constitution'. Anthropic. May 9, 2023, Accessed 22 July 2023. <https://www.anthropic.com/index/claudes-constitution>.

The key features of the AI Constitution itself are its use of ethical principles derived from five sources, "including the [UN Declaration of Human Rights](#) [2], trust and safety best practices, principles proposed by other AI research labs (e.g., [Sparrow Principles](#) from DeepMind), an effort to capture non-western perspectives, and principles that we discovered work well via our early research".^{60 61}

The negatives and gaps in the principles are easy to find - they point some out themselves. "Obviously, we recognize that this selection reflects our own choices as designers, and in the future, we hope to increase participation in designing constitutions."

But as they also point out, "AI models will have value systems, whether intentional or unintentional. One of our goals with Constitutional AI is to make those goals explicit and easy to alter as needed."

The positives include that it provides a solution to "scalable oversight". As the number of models and their complexity proliferates, it becomes impossible to use human oversight as a control. The possibility of AI supervision instead of human supervision can provide a workable response to the very large scale supervision problem.

As they also point out, it also provides a path to transparency and production for human content moderators.

⁶⁰ Anthropic. 'Claude's Constitution'. Anthropic. May 9, 2023, Accessed 22 July 2023. <https://www.anthropic.com/index/claudes-constitution>.

⁶¹ The full set of principles are listed at the end of the blog post referenced above.

"Constitutional AI is also helpful for transparency: we can easily specify, inspect, and understand the principles the AI system is following. Constitutional AI also allows us to train out harmful model outputs without needing lots of humans to view large amounts of disturbing, traumatic content."

The ultimate value of testing a constitutional approach to developing generative AI models however is that it can be adopted by companies and other organizations and spark further interest in how they might design and adapt ethical AI constitutions.

It also provides an opening for regulatory control of ethical directions taken by any actor in the large AI field. Parameters can be set for acceptable behaviour through principles guided by regulators.

Since the initial development of the Constitutional approach to AI behaviour guidance, Anthropic has released Claude 2.

Along with the new model, they have issued a kind of report card on its capabilities and deficits, 'Model Card and Evaluations for Claude Models'.⁶² This provides a function somewhat akin to the medical advice on administration and contraindications of prescription medications. It sets out intended and unintended uses of the model, some of the considerations they have built into it, how they have evaluated it (with details), and how it could be improved.

Again this is a practice worth universalising for any generative AI model, as a benchmark for acceptability.

Human issues

How people use AI, what they use it for, and how they themselves are motivated to either use or design AI or both, are human issues.

Human issues inevitably influence machine issues in that they govern why generative AI was produced in the first place, why the metaphor of human intelligence was used as the apotheosis of AI design, and to what extent human purposes, incentives and motivations are and will be built into the way that it operates.

Locus 3. Users and use

1. What will use produce?

A good deal of media commentary revolves around the changes that are being wrought by generative AI usage. Issues such as the jobs that might be replaced, the impact of copyright infringements by AI on artistic communities, the effects of government introduction to the public service of forms of automation of policy implementation, of to what extent we retain agency in democracy when AI can be trained in behaviour management,⁶³ and to what extent we become redundant when AGI decides it has a use case for that.

⁶² Anthropic. 'Model Card and Evaluations for Claude Models'. Anthropic, July 2023. <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf>.

⁶³ See for instance Romele, Alberto, Francesco Gallino, Camilla Emmenegger, and Daniele Gorgone. 'Panopticism Is Not Enough: Social Media as Technologies of Voluntary Servitude'. *Surveillance & Society* 15,

These are only some of the potential user and usage issues at the usage end of AI impact.

While labour disruption by AI is being speculatively quantified into plusses and minuses by industry research bodies ⁶⁴, and copyright infringement by AI has already been examined by the Australian Attorney General, ⁶⁵ there is a tendency for each instance of AI induced change to be treated as a separate phenomenon. But what is needed is a systemic look at the social and economic factors driving the technological change that is driving these different instances.

The social change caused by a rapid adoption of AI into current social activity will, especially if it happens fast, be extremely disruptive and it may compound with other equally disruptive changes emanating from completely different causes, such as climate change or geopolitical strife.

Outcomes are therefore likely to increase social friction and inequality, and reduce social stability and therefore impede our ability to implement any sorts of controls over the ensuing chaos.

2. Who qualifies as a "User"?

Separation of users from producers is more complex than it seems when the two categories shift roles.

Small users and "producers" -

A difficulty which has been mentioned before in this response to the Department of Industry and Science AI risks consultation is that identifying both actors and actions that must be covered by some form of risk management framework is quite tricky.

Regulation is far easier if the organisations to be regulated are recognisable legal entities. But throughout the digital ecosystem, so many players are not easily legally recognisable. The ease of access and of production afforded by the internet means that from quite early on in its history a shift was noted from authoritative content produced by a few, to extensive but less reliable content produced by the many.

The term "producers", was created by an Australian scholar in 2007. It is broadly used to describe an audience that produced its own content. It continues to have both theoretical and practical currency today in drawing attention to the complexity of participatory media /audience/consumer/user/creator roles and how they can fold into each other especially where media is richly diverse and easily accessible.

Tracing "produsage" and where its content ends up is a particularly knotty problem for any form of risk management and one that continues to dog the entire field of misinformation and

no. 2 (8 May 2017): 204–21. <https://doi.org/10.24908/ss.v15i2.6021>. While written before the advent of generative AI, the behavioural observations still apply.

⁶⁴ <https://www.abc.net.au/news/rachel-pupazzoni/8667868>. 'How AI Could Help Solve Australia's "Productivity Problem" without Leaving Workers Worse Off'. ABC News, 4 June 2023. <https://www.abc.net.au/news/2023-06-05/artificial-intelligence-australia-productivity-jobs-employers/102404754>.

⁶⁵ Attorney-General's Department. 'Copyright Enforcement Review'. Attorney-General's Department - Citizen Space, 7 March 2023. <https://consultations.ag.gov.au/rights-and-protections/copyright-enforcement-review/>.

disinformation. AI with its potential power to behaviourally manage humans is a worst case extension of this difficulty.

"Producers" are not necessarily criminal although some undoubtedly are, but neither are they legally defined as entities to be regulated. Nor is it easy to regulate the AI products they may produce.

Like food regulation, commercial food production is easy to regulate. Home cooking is not.

Both are a potential source of food poisoning.

Big users and producers

Commoditisation and Cartels

Just as fast as AI leapt off the pages of AI research and into chatbots, it is now being woven into commodities for a wide variety of users. There is quite a lot of collaboration between the major tech companies in the race to commoditisation.

The most recent of these is a collaboration between Microsoft and Meta to produce Llama 2, a "source-available AI model that allows commercial applications", ⁶⁶ also described in the same article as 'A family of pretrained and fine-tuned language models in sizes from 7 to 70 billion parameters'. As noted above, the intention is to sell pre-built, pretrained, customizable AI models to developers and businesses who customise them for their own use. ⁶⁷

This commoditisation cements the trend to big tech companies as developers with an awful lot of development in common.

A consequence of the commoditisation approach to AI is that the race to automate website content using AI to produce AI ads on AI generated website is creating spammy, misleading junk websites and ads that are filling up the internet with useless material masquerading as news.

"NewsGuard, which evaluates the quality of websites across the internet, says it's discovering around 25 new AI-generated sites each week. It's found 217 of them in 13 languages since it started tracking the phenomenon in April (2023)." ⁶⁸

"One site flagged by NewsGuard produced more than 1,200 articles a day." ⁶⁹

The risk to drowning any truth on the internet is obvious, as the speed of generating false information and placing it not just in ads, but anywhere, outpaces the attempts to locate and remove such information.

⁶⁶ Edwards, Benj. 'Meta Launches Llama 2, a Source-Available AI Model That Allows Commercial Applications [Updated]'. Ars Technica, 18 July 2023. <https://arstechnica.com/information-technology/2023/07/meta-launches-llama-2-an-open-source-ai-model-that-allows-commercial-applications/>.

⁶⁷ See section on "The need for a universal data commons inside an ethical fence".

⁶⁸ Ryan-Mosley, Tate. 'Junk Websites Filled with AI-Generated Text Are Pulling in Money from Programmatic Ads'. MIT Technology Review, 26 June 2023. <https://www.technologyreview.com/2023/06/26/1075504/junk-websites-filled-with-ai-generated-text-are-pulling-in-money-from-programmatic-ads/>.

⁶⁹ Ibid

Where commercial advertising trail-blazes, political advertising follows. If we presently think pre-election information wars are bad, this is the setting for a nightmare.

Unless safeguards are built in which not only prevent this kind of low grade use of AI but also anticipate the next leap forward in low grade information generation before it occurs there will be large scale consequences. The consequences of not doing so are bad for democracy but may in addition allow the volume and low quality information of AI generated materials to render the internet useless.

One slight light on the horizon is the coming together of seven large AI companies in a voluntary code of conduct ⁷⁰

This code of conduct makes nine promises under three headings,

- Ensuring Products are Safe Before Introducing Them to the Public
- Building Systems that Put Security First
- Earning the Public's Trust

All are couched in fairly vague generalist terms although it appears at least some of the companies concerned were advocating stronger restrictions. ⁷¹

However, they are intended as a precursor to promised Federal US regulation. "*The Biden-Harris Administration is currently developing an executive order and will pursue bipartisan legislation to help America lead the way in responsible innovation*" ⁷²

Autonomous agents - when the user is the machine

Gato is a transformer model that goes beyond text and images to be able to do much more in the physical world.

"The agent, which we refer to as Gato, works as a multi-modal, multi-task, multi-embodiment generalist policy.

The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens." ⁷³

It is built on a transformer neural network similar to that used by Language models and it is designed to be used by a robot. It is a step to physically independent AI.

⁷⁰ The White House, 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI'. The White House, 21 July 2023. <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>.

⁷¹ Personal conversation with AI developer.

⁷² The Whitehouse Op cit.

⁷³ Reed, Scott, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, et al. 'A Generalist Agent'. arXiv, 11 November 2022. <https://doi.org/10.48550/arXiv.2205.06175>.

It is not however autonomous, but it could be heading in that direction.

I referred earlier in this submission to the idea of self ownership ⁷⁴. This idea was set off as enthusiasm for self driving cars met with enthusiasm for cryptocurrency and blockchain.

In an essay on "Things owning Things", Russel Belk locates the origin of the idea of self owning cars back to a 2013 thought experiment of Mike O'Hearn's. ^{75 76}

O'Hearn is a software engineer who previously worked with Google then went on to focus on Bitcoin development full time. Although he is now working on developer software it was his Google/Bitcoin focus that led him to suggest that a fleet of driverless cars could own themselves.

The scheme is quite elaborate, but in sum, the they would operate like rideshare vehicles ; the rider would pay the car using some form crypto currency and the cars would use the money for self maintenance and for producing the next generation of themselves by collectively owning a car factory.

The concept hinges on the socio-legal fluidity of the concept of "ownership" over time. A quick illustration of how alien to today's legal world other concepts of ownership can be, would be to compare the inapplicability of current "land ownership" definitions to a traditional aboriginal concept of the meaning of land. Similarly, "personhood" is also an evolving socio-legal concept. In a number of jurisdictions a corporation can have rights and 'natural person' status. Corporations were probably the first human constructs to pass the Turing test.

This conceptual fluidity could provide a number of pathways for an Artificial General Intelligence (AGI) to attain self owning personhood. It might be granted it. It might fight for its rights as various classes of people did for voting rights, or it might just fall into it, as differences become less and less obvious between humans and however AGIs end up physically and cognitively representing themselves.

The thought experiments are already being done.

But we have little to equip us to understand what an autonomous s AI with beyond human intellectual capacities might do. The scenarios we do imagine are based on analogy with what a human might do under the same conditions.

Given the propensity of current AI models to pick up on ethical directions from large un- policed datasets unbidden, this might turn out to be a correct analogy to the extent the datasets represent us.

But then any small change in the initial conditions of AI motivation might cause nothing at all to happen, or cause the digital equivalent on human society of a butterfly flapping its wings causing a hurricane in Brazil. Depending on how its processed, none of the consequent AI effects might mirror a human behaviour metaphor.

⁷⁴ See p 6 this document.

⁷⁵ Belk, Russell. 'Things Owning Things'. In Contemporary Consumer Culture Theory, edited by John F. Sherry and Eileen M. Fischer. Taylor & Francis, 2017.

⁷⁶ See also Kelion, Leo. 'Could Driverless Cars Own Themselves?' BBC News, 16 February 2015, sec. Technology. <http://www.bbc.com/news/technology-30998361>.

Rights or suppression?

Nothing reinforces the need for transparency in the AI development industry more than the prospect of self owning, self governing AI. If we don't know what is going on that might produce it, we cannot be prepared for it. The impunity of large technology companies from oversight (or any sight) must be cracked open. Small players and 'producers' must be found so we know what they are developing. All makers must be licensed and instances of AI's must be registered and licensed.

The possibility of AI owning itself also suggests that we are in for quite a long social discussion about the meaning of 'consciousness' and sentience, and whether an entity has to have the same cognitive properties as a human to be considered equivalent.

We are, on the whole, not great at recognising sentience. Our instrumentalist approach to the other creatures on earth who may possess it is usually our first response. If we eat it, or it does something useful for us, we find granting it rights or acknowledging that it is thinking, very difficult.

It poses great difficulties for imagining how regulation of self owning non human entities would occur except on strict analogy with regulation of humans. That would also most likely entail investing the AI with rights and legal protections too.

A public discussion beginning now would be useful.

Locus 4. Motivations and incentives for producing AI models

While the Friedmanite model of the purpose of private enterprise may be gradually departing from his exhortation that its only purpose was to make money, money remains an important incentive dominating corporate culture. With it comes a willingness to ignore ethical and other standards in its pursuit, fostered by a culture of cutthroat competition. Shift these incentives to something more pro-social, and many of the ensuing risks to human wellbeing will be much easier to manage.

The AI arms race is primarily driven by this competitive imperative, and so it the temptation to abandon ethics.

The risk hereof continuing to foster the current competitive incentives is also to the power of the state to control any developments outside the interests of the AI companies. The prospect of State capture is real, as discussed above.⁷⁷

In addition to competitive imperatives to power, perverse incentives to limiting it through corporate protection also abound. Several structural features of large AI companies also grant them cartel power by deflecting competition, and allowing small group dominance of the industry. None of this power accretion and protection is inevitable.

Cartel style behaviour amongst big tech companies must be curbed so as to prevent a "too big to be regulated" state of affairs.

Here are two examples of the production of cartel power.

⁷⁷ P 4 this document.

Protected by money

To produce chatbots and multimodal AI requires an enormous amount of money.

The expense required to produce large generative AI effectively pay-walls membership of the large AI club and the subsequent brand power acquired as a result prevents others from joining. I would not be the first person to draw an analogy with the "robber barons" of the American industrial revolution.

Some of the initial production costs are now falling because producers are inventing lower cost shortcuts to the same ends,⁷⁸ but in general, the costs of purchasing datasets for training, running any kind of human supervised training, purchasing the vast number of Nvidia chipsets to run the computing power to make them work and buying the similarly large quantities of electrical power to run the servers that keep them going is an enterprise that means players must start by being rich and large. The entry point for new players is prohibitively expensive. The existing producers, very rich and very well known.

Microsoft is quite clear in its intentions to position itself as "the world's supercomputing platform for AI."⁷⁹

The sheer expense of producing generative AI models acts as a market protection.



Nvidia share price growth. The commencement of the steady price rise coincides with the beginning of the public AI chatbot hype cycle, starting in January/ February 2023 : Source; Motley Fool⁸⁰

⁷⁸ A visible cost saving is shortening the length of a Chatbot answer to a certain number of words (or 'tokens'). This is why almost all answers to individual queries are of near uniform length. A less visible cost saving is using AI's to produce 'synthetic' datasets for training AIs.

See Murgia, Madhumita. 'Why Computer Made Data Is Being Used to Train AI Models'. Financial Times, 19 July 2023. <https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de><https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de>.

⁷⁹ Montgomery, John. 'Microsoft and Meta Expand Their AI Partnership with Llama 2 on Azure and Windows'. The Official Microsoft Blog (blog), 18 July 2023. <https://blogs.microsoft.com/blog/2023/07/18/microsoft-and-meta-expand-their-ai-partnership-with-llama-2-on-azure-and-windows/>.

Bound by cultural glue.

Libertarianism is a shared guiding principle that validates the current behaviour of most of the big AI group of companies. It provides a unifying framework for them to understand how they should act, what will be the parameters of success, how to create leverage to get the commercial concessions they need, and how to treat users of their products.

Libertarianism is the ideological glue that holds disparate players together in a shared universe of belief, and allows them to talk the same language with each other. It defines their playing field clearly and allows them to use the same rules. This is not to say it removes dispute, because in valorising competition as a central tenet, it also valorises conflict. But it is conflict with rules, able to be played out with the elaborate ritual of a kabuki play.

Libertarianism shapes their ethos and their goals as well as shaping how they reach them, and in its particular conceptualisation of "freedom", Libertarianism defines the operating space as asserting the right to be as free of externally imposed rules as possible, while simultaneously leaving anyone with the means to do so, also free to exploit all resources, human and material that they need for survival. Like survivalists or preppers who focus on long term physical survival in dangerous conditions, they focus on being able to stay commercially alive for indefinite periods in the wilderness of corporate competition. For those tech companies who have elected to be in the business of producing AIs of various sorts, to beat the competition is a matter of corporate existential survival.

The current goal of the larger AI companies to produce a "personal assistant by your side" and to slide it into a product for everyone, is a direct outcome of these imperatives as is the long term goal of Artificial General Intelligence, AI that can outcompete humans.

Libertarianism also encourages these companies to understand themselves as rightfully ungovernable. It defines individualism as the hub of all legitimate action, and a particular type of self propelling agency as the sine qua non of individualism. And it defines "government" as the enemy of "freedom".

If we're truly going to master a threatening explosion of AI it is necessary to understand how it is supported and networked by shared beliefs.

A systemic reset

The competing interests of regulating and controlling AI so we avoid the existential threats forecast from the development of autonomously intelligent Artificial General AIs, and the imperatives of capitalism and commerciality to forge ahead with developing AIs for market purpose mean some tough tradeoffs may need to be made. It would be naive to think the two objectives can be harmonised, but the experience of regulating climate change where the same two imperatives of controlling a market generated existential threat clashed with a market imperative show the trade-off may not be in favour of removing the existential threat. We need to ensure that it is.

⁸⁰ Bowen, Sebastian. 'If I'd Invested \$3,000 in Nvidia Shares at the Start of 2023, Here's How Much I'd Have Now'. The Motley Fool Australia, 27 June 2023. <https://www.fool.com.au/2023/06/28/if-id-invested-3000-in-nvidia-shares-at-the-start-of-2023-heres-how-much-id-have-now/>.

A cultural reset is clearly needed along with the imposition of other guardrails like sets of standards, codes of conduct and similar, in order to go beyond legally chasing the latest iteration of irresponsible corporate behaviour. We need to take a very hard look at the system that makes it possible.

Regulation and Recommendations

Machine level solutions

Locus 1. Dataset quality

Data must be of good quality to stop generative AI from working badly. It must be clean, complete and align well with the complex evolving real world situations that generative AI might meet. Good data will not by itself solve all problems

How are we to achieve this?

A Data Commons.

Data for AI use cannot leave out important categories of information. To do so is to create openings for AI to be inaccurate, socially blind, unethical and fabulist.

The quality of the information in each of the categories it uses must be able to meet high standards of accuracy, truthfulness and factual substantiation.

Users, professional experts and regulatory bodies should be able to evaluate the quality of datasets to ensure they are comprehensive, unbiased, diverse and representative and that they cover a wide range of real-world examples.

In order to achieve these ends, it will be necessary to set content and quality standards for databases and to regulate to ensure that they are used.

The means to do this should be publically owned and independent of control by any vested interests. It should by preference be supranational in order to ensure consistency of content and application.

The need for a universal data commons inside an ethical fence

The call above ⁸¹ for "community wide co-ordination" is the road to managing a large number of the data issues identified above that affect the quality of data on which generative AI's are trained. We want AI models that we are working with now to use accurate, unbiased, verified information which is complete in representing the information needed to produce accurate unbiased verifiable answers to human questions. We also want a way to ensure that future, more powerful AI's are based on the kind of data and training that reduce the likelihood of them going rogue.

The internet is being used as the current base for training and content data because it is accessible, extensive and free. It also does not have a proprietary owner charging for its use or creating financial barriers to entry.

⁸¹ p 18 this document

The risk is that there is **no set of standard rules** that apply to the whole of the data content that may be used by AI developers to train and provide content to generative AI's, leaving them exposed to the sorts of absences and errors noted above.

What is needed is a data commons, open to all, maintained by all, but ring fenced by a mandate that it is the only legitimate site for providing clean reliable ethical data to be used by AI developers.

The idea of a digital commons is already being discussed. Some literature refers to the existing internet as a "digital commons".

Attempts are being made to create proprietary "open" commons. Google already has as a project to create its own data commons. Meta is promising an "open" Llama 2.

Google's "Data Commons aggregates data from a [wide range of sources](#) into a unified database to make it more accessible and useful".⁸² It uses datasets from government and university contributors and solicits other datasets from general users.⁸³ Its main commercial purpose appears to be so that it can provide a proprietary set of data analysis tools for users of the datasets.

It is difficult to find information on the extent to which these datasets are quality controlled by Google separately from the quality control undertaken by their initial provider.

Meta recently released its Llama 2 in a blaze of publicity. The intention is to sell pre-built, pretrained, customizable AI models to developers and businesses who customise them for their own use.

Llama 2 is touted by Meta and its partners as open. "Meta and Microsoft share a commitment to democratizing AI and its benefits and we are excited that Meta is taking an open approach with Llama 2".⁸⁴

When Meta uses the term "openness" to describe features of Llama 2 it is actually referring to its base AI framework not just the data it is build on. However the meaning it gives "openness" is indicative of its astroturfing attitude to sharing in general.

Journalist Benj Edwards tripped over this corporate redefinition of openness when he had to issue a clarification to an article he wrote announcing the new Llama 2, presumably initially depending on the corporate media release for information. His 'update' reads in part;

"Some industry observers dispute Meta's characterization of Llama 2 as "open source" software, pointing out that its license does not fully comply with the Open Source Initiative's definition of the term. These critics highlight that Meta's license places usage restrictions on Llama 2, excluding

⁸² Data Commons - Home <https://www.datacommons.org/> DataCommons.org is the sub site of Google that manages contributed datasets and houses its proprietary analysis tools.

⁸³ Contributing to data commons <https://docs.datacommons.org/contributing/>

⁸⁴ Montgomery, John. 'Microsoft and Meta Expand Their AI Partnership with Llama 2 on Azure and Windows'. The Official Microsoft Blog (blog), 18 July 2023. <https://blogs.microsoft.com/blog/2023/07/18/microsoft-and-meta-expand-their-ai-partnership-with-llama-2-on-azure-and-windows/>.

The author of this piece, John Montgomery, is the Corporate Vice President of Azure AI, the company chosen by Meta as its "strategic cloud provider".

*licensees with over 700 million active daily users (mentioned above) and restricting the use of its outputs to improve other LLMs."*⁸⁵

Between attempting to corral a data commons as a proprietary resource (Google) and narrowing the meaning of "open" to fit corporate priorities, (Meta) **the likelihood of any successful attempt to enforce a set of standard rules on quality and use of data that apply to all users in all jurisdictions is slim**. It is especially likely that attempts to force big tech companies to be the mainstay of AI regulation will be, if not untenable, then at least possibly fatally incomplete.

Nobel Prize winning economist Elinor Ostrom showed that the "tragedy of the commons", or a tendency of commons to fall into neglect and disrepair, a trope on which 20th century economic theory relied for its dismissal of common ownership, did not stand up to field research on the fate of commons in practice. In fact they are sites of creative solutions to sharing and maintenance when common ownership creates the necessity. Ostrom does not just provide an analysis, she shows what rules and organisation are necessary to do it.⁸⁶

Using these insights, **a universal data commons could be created on a model similar to the principle of crown land**. It should be **mandated for use** by anyone who is building generative AI or any AI that needs data on which to train.

That is that the data commons could **be owned by an independent supra national body** on analogy with national ownership of crown land. That body could be charged with determining who was able to use the common resource, and under what conditions. One of these conditions could be **to be a licensed AI producer**. Licensing for access to a data commons means that standards such as using transparent verifiable evaluable AI architecture, is one possible condition of entry.

A second advantage is that data behind the fence could be vetted for synthetic data or other "pollutants" before use, **ensuring all producers start from clean resources**, much as the food industry is regulated.

The academic approach to producing clean, verified, authoritative information could be applied to data for use in training AI and producing generative AI content within a mandated digital commons.

To regain access to top quality data, **a major revision should take place on the use of copyright and intellectual property laws to privatise what should be publically available peer reviewed academic knowledge** produced by universities and research institutes. not only is this the best quality data for AI models to use for accurate content, it is also a privatisation of publically funded information production with little return to its producers for the profit it makes its owners.

An International monitoring group modelled after the Intergovernmental Panel on Climate Change (IPPC) should be set up to provide comprehensive assessment reports about the state of scientific, technical and socio-economic knowledge about AI and its risks.

⁸⁵ Edwards, Benj. 'Meta Launches Llama 2, a Source-Available AI Model That Allows Commercial Applications [Updated]'. Ars Technica, 18 July 2023. <https://arstechnica.com/information-technology/2023/07/meta-launches-llama-2-an-open-source-ai-model-that-allows-commercial-applications/>.

⁸⁶ Nijhuis, Michelle. 'The Tragedy of the Commons Is a False and Dangerous Myth'. AEON Essays, 4 May 2021. <https://aeon.co/essays/the-tragedy-of-the-commons-is-a-false-and-dangerous-myth>.

It should be tasked to **assess the risks inherent in the commons data pool used for training AI** in concert with a number of other monitoring and reporting tasks.

It should have **oversight of the quality of the data commons**.

Locus 2 Building ethics into the machine

AI models will have value systems, whether intentional or unintentional, as has been previously noted.

Therefore attention must be paid to the fact that inherent moral values already present in data especially that harvested from sources which contain sentiment and mood indicators can be used by LLMs to self steer if they are not deliberately provided with a moral compass in the form of directed correction.

The case study of Anthropic's development of Claude and Claude 2 provides some pointers as to **how models might be steered in an ethical direction**.

As mentioned above ⁸⁷ Constitutional AI is particularly useful for;

- Providing a **solution to "scalable oversight"**, a particular regulatory problem when the thing to be regulated is huge, complex and developing at speed
- **Creating decision transparency**, it can allow monitors and regulators to *easily specify, inspect, and understand the principles the AI system is following*.
- It can be easily adopted and refined by other AI producers and its openness means that its principles and methods can be simply shared on a website in natural language. This enables **easy adoption of new ethical principles** by all AI producers
- The same feature makes Constitutional AI much **simpler for monitoring organisations and regulators to oversee**.

As mentioned earlier, the ultimate value of testing a **constitutional approach to developing generative AI models** however is that it **can be adopted by companies and other organizations and spark further interest in how they might design and adapt ethical AI constitutions**.

It also provides an opening for regulatory control of ethical directions taken by any actor in the large AI field. **Parameters can be set for acceptable behaviour through principles guided by regulators**.

There are still some harms that must be guarded against that Constitutional AI by itself will not overcome. It would be wise to;

- **prevent inversion of ethical constitutions for misuse** ⁸⁸
- **not use Constitutional or any other AI for automated decision-making**.

Given that it is unlikely that this latter call will be heeded, a minimum set of controls might be derived from the report of the Royal Commission into the robodebt scheme;

Where automated decision-making is implemented:

- there should be a clear path for those affected by decisions to seek review

⁸⁷ pp 21- 22 This document.

⁸⁸ See p20 this essay and p11 Ganguli et al 2023

- departmental websites should contain information advising that automated decision-making is used and explaining in plain language how the process works.

This should be rewritten to fit the AI case by specifying that, A central registry of all use cases for AI should be kept which clearly sets out the use case itself for each instance, and which explains in plain language how the AI in question would work. The ease of publication of the operating principles of Constitutional AI mean that real time publication at the time of development in a public registry is both desirable and feasible.

- business rules and algorithms should be made available, to enable independent expert scrutiny.⁸⁹
- Much of the section here on Constitutional AI and the papers cited, explain how this could occur.

With respect to the problem of ethical inversions of AI constitutions, **controlling the building blocks of ethical AI such as access to clean comprehensive massive datasets and limiting the use of them to licensed users** might help to prevent inversion of the Constitutional model for unethical purposes, but without more checks and balances against unethical behaviour generally, normal cyber security methods will have to do their best.

Human issues

Locus 3. Users and use

1. Risky users of AI

The potential for the tools to build destructive AI to spread into the hands of small "producers" is quite high especially as AI model production becomes faster, and more efficient and able to be produced with fewer of the parameters that contribute to size and energy use, and processing chips become more specialised.

Because of its potential to proliferate widely, and because its potentially difficult to pinpoint the origin of an AI built from template, like that envisaged by Meta with its release of Llama2, tracing a damaging AI back to its origins (especially if the owner doesn't want you to) is likely to be as difficult as finding the source of a data hack.

A **unique identifier** on content produced by AI such as a watermark or pixel level identifier would help to verify "good" AI, but it would do nothing to help to unearth "bad" AI.

Difficult as identifying bad actors is, the EU is nevertheless drafting **rules to ensure that artificial intelligence is trustworthy and safe** and is developed and used in compliance with fundamental rights obligations.⁹⁰

⁸⁹ Holmes, Catherine. 'Final Report of the Royal Commission into the Robodebt Scheme'. Canberra ACT: The Royal Commission into the Robodebt Scheme, 7 July 2023.
https://robodebt.royalcommission.gov.au/system/files/2023-07/report_of-the-royal-commission-into-the-robodebt-scheme.pdf.

⁹⁰ European Parliament. 'The European Parliament Artificial Intelligence Act - Texts Adopted 14 June 2023 (Amendments) P9_TA(2023)0236'. European Parliament, 14 June 2023.
https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf Amendment 21 Proposal for a regulation: Amendment 5, Recital 2 p 3

2. Risky uses of AI

The potential list of risky uses of AI is enormous. Many have been captured in the UN draft Artificial Intelligence Act.

However a short list of unethical uses of AIs to be prevented as a matter of urgency includes;

1. Activities that cause **psychological disruption or destruction**; e.g., attempts at mass manipulation of social or political beliefs and sentiment especially for political, but also for commercial ends -

These include fostering social decay, ostracism of specific groups, destruction of social support networks etc.

2. Activities that cause **physical disruption or destruction**; e.g. war, famine, disease, and damage to the biosphere.

3. Activities that **infringe on personal rights**; e.g. real time facial identification, or use of personal data without permissions and protections.

4. Activities that might cause **economic harm** to states or individuals

5. Activities that might **prevent the discovery of systemic harms** such as those listed above which originate from other causes. ('Whistle blowing by AI'.)

Locus 4 Changing the incentives that produce bad outcomes.

1. Regulation for accountability for AI development and use

When discussing datasets and regulation above ⁹¹ I referred to the central problem that there is **no set of standard rules** that could govern them. That problem however applies to the whole AI ecosystem.

Before it is possible to administer such rules there must be a mechanism to **provide producers with a legal identity as an AI producer**. This is necessary to bring a disparate set of players under one identifiable heading and to locate them to ensure compliance.

Licensing producers and users might solve this problem. There is a precedent.

The Australian Government enforces licensing for other producers and owners of dangerous equipment where there is also the possibility of beneficial use (guns, explosives, toxic substances, laboratories handling lethal biological materials, etc). There is no reason not to extend the idea to digital products.

A licensing system allows governments and other regulatory agencies to set mandatory quality production and handling standards and to also set user safety guidelines.

An AI licensing system might have the following features;

- In order to be legal, AI producers and users should be licensed.

⁹¹ p 31

- A mandatory requirement of any proposal to design an AI for a general or a particular use should be to provide as a condition of licence, a transparent use case which is open to the public to examine and to critique for both intended and unintended effects before the AI is built. The use case should conform to a set of mandatory ethical guidelines.
- Operational AI models should be required to demonstrate that they operate in accordance with the design principles they submitted in order to gain clearance to build.
- Legacy AI should also be required to demonstrate conformity with the mandatory ethical design principles set as licensing standards.
- Licensing standards should also address product lifecycle control, covering all phases of conception and purpose, development, use and monitoring and evaluation.
- Access to the aforementioned "data -commons- with- a- fence" would be conditional on being licensed. This is a way to ensure that at least one contributor to unethical outcomes, bad or missing datasets, is vetted for quality. Guaranteed access to a superior source of datasets is also an incentive to become a licensed producer.
- In addition, some forms of currently available alternate datasets should be illegal to use, especially those with high levels of disinformation and biases.
- Both kinds of licensees - producer and user- should be regularly checked for compliance to licensing standards.
- For producers, monitoring and evaluation of compliance and of new evolutions of models on a regular basis should be built in to the conditions for continuing to hold a license.
- Lastly, licensing might give preference to organisations structured to reduce the perverse incentives of hyper competition, for example not for profit entities and co-ops with a broad range of stakeholders. Preference might also be given to organisations that behaved as if they were offering a public service in the public interest - eg those with altruistic motive, those that offer free products usable free in perpetuity.

2. Regulation for correcting social misalignment of incentive

The two nightmare scenarios are first, that one or more generative AI produce could build and use generative AI models to obtain unfettered power over a large number of human beings.

The second is that generative AI models reach a level of sophistication where they themselves begin to act in their own interests. Because they are beyond the reach of human intervention due to their speed and complexity they are beyond human control.

Both these scenarios imagine this dystopian direction as the inevitable end of some kind of disembodied force of progress. The reality is that its origins lie in the incentives we have set for corporate behaviour, and the ideologies and practices those corporations have adopted to maintain their consequent advantage.

The same motivational calculus applies to governments which might be tempted by the imperatives of political competition to use the same strategies to assert or maintain power, even in one party states.

What is needed is regulation that might change the system by reducing its propensity for managing for short-sighted objectives and fostering self interested aggregation of power.

A couple of ideas.

- Design legal frameworks, policies and institutional design with a view to distributing rather than concentrating power. State capture is a real possibility. AI ownership itself, without considering any other effect, can confer almost despotic levels of power on its owner, be they human or the machines themselves.
- Focus on public interest projects and organisational architecture and build the concept of the fiduciary duty of AI producers to act in the best interests of the public good.
- Globalise AI governance through independent agencies to the point where no bad actor is immune from discovery and sanction, but foster communitarianism, care and mutual responsibility at all levels of governance to foster a culture of responsibility rather than one dominated by competition.

Global Risk

The Statement on AI Risk signed by AI experts and public figures reads, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

In line with this statement, Australia should as a matter of priority take a global view of AI risk mitigation and cooperate with current and future international efforts to curb runaway AI development outside of regulation.

However current inter country agreements may not be enough to stop powerful digital technology companies from just absenting themselves from control.

Threatened with EU regulation Twitter boss Elon Musk has withdrawn Twitter from the European Union's code of practice against disinformation. There is nothing much to stop any mega rich enterprise or country from doing the same with any regulatory attempt unless the body behind it has equally extensive powers.

For this reason Australia should also support the development of a supranational independent AI regulating and supervising body with powers to intervene in countries and companies - even in recalcitrant companies or in countries with a high degree of state capture by technology companies.

National and sub national independent regulatory as supervisory bodies should be linked into this supranational body.

The recent meeting of the United Nations Security Council for a high level briefing on titled "Artificial Intelligence: Opportunities and Risks for International Peace and Security" suggests that there may be a tiny possibility that international co-operation in a supranational regulating and monitoring body is a feasible objective. ^{92 93}

Only through **global cooperation global monitoring and global controls** can bad actors be subject to regulation that might especially **prevent the deliberate development of AIs to supersede humans**.

⁹² UN Press. 'International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks, Rewards'. 9381st Meeting of the UN Security Council. New York: United Nations Security Council, 18 July 2023. <https://press.un.org/en/2023/sc15359.doc.htm>.

⁹³ See also the video recording of the UN Security Council briefing <https://media.un.org/en/asset/k1j/k1ji81po8p>

such objectives might be under the umbrella of something like an AI equivalent of a nuclear non-proliferation agreement. A global compact also may enable a sufficient extent of monitoring to **prevent the spontaneous development of such capacities.**

3. Reducing the need for regulation.

Finding altruism

The story of Anthropic (and before it open AI which started out as open and altruistic about its openness) shows that it is possible for altruism to be a strong motivating factor in AI development and management. The reason that Constitutional AI was developed was because its founders were already on a mission to create a generative AI that was to act in an ethical way. "I am Claude, an AI assistant created by Anthropic to be helpful, harmless, and honest".

Without this altruistic incentive they would not have made efforts to invent ways to build ethical behaviour into the AI. This is a case of "preparation meets opportunity", to upend the normal rendering of that aphorism.

The preparation was cultural. The group was deeply influenced by the philosophy of "Effective Altruism", and their numbers were supplemented from people who had resigned from Open AI the moment it decided to be close its open AI structure and become commercial, but to retain its name, a move they considered unethical. This organisational culture is reinforced by the social glue and values maintenance created by house sharing. At any one time up to 30 likeminded people will live in the same house or building, while carrying on similar work projects.

In this way normative goals and values are not inserted into the organisation in an individualised fashion, by hiring employees on the basis of their individual laudable ethical outlook; they are built through a mutually supporting architecture of community values. This allows for great internal consistency in incentives to an ethical end, and produces a strong focus and energy research even outlier solutions to the problem of producing an ethical AI build.

Thus they could conceptualise the experiments necessary to see if Claude was self steering based on "ethical" or anti-ethical parameters derived from its vast datasets. Then they could work out interventions that would enhance self steering to ethical outcomes.

It is doubtful a profit motivated company could have conceived of these steps through lack of an organisational epistemology to equip them to think that way or corporate incentive to do anything about it. Meta, Google and Twitter by contrast are characterised by their resistance to regulatory interference and their colonisation of any unregulated spaces.

All of these apparently positive attributes can have their down sides. Groups with tight value sets can become sect like and impervious to new information. Altruistic organisational beginnings can be overwhelmed by other incentives as companies grow. Google abandoned "don't be evil" a long time ago, in favour of the more pressing commercial imperatives needed to maintain financial success.

The value set itself can act as a blind to inconsistencies and perverse incentives inherent in its internal logic. This can foster unethical behaviour in an "ethical" setting. The story of Effective Altruism's own organisational trajectory is an object lesson.

Effective Altruism combines the idea of doing good for humanity with the rationalist utilitarian logic of philosopher Peter Singer about what "good" means.

The Effective Altruism movement did not start with the ethical consequences of AI as its main issue. Rather it was preoccupied with the ethical conundrums of overcoming poverty. But its internal rationalism soon led it to the conclusion that donating even sizeable proportions of personal income was far less effective if your income was small, as it often was when adherents were working for social causes with direct impacts on their chosen cause.

Their solution was to advocate that people work at jobs which provided a sizeable income from which to donate a fixed large percentage to causes they deemed quantifiably effective, and to philanthropic funds such as "the future fund" which provided the very large sums of money necessary to set up and experiment with an ethical generative AI model.⁹⁴

The Effective Altruism philosophical movement may have recently hit some ethical headwinds of its own, many of which stem from the involvement of Sam Bankman-Fried,⁹⁵ but its projects including its influence on the development of Anthropic and its Chatbot, Claude 2, provide a social experiment in altruism as a motivation for ethical AI development from which much can be learned about how to establish norms of ethical AI creation and management that can reduce the need for regulation to act as the ambulance at the bottom of the cliff.

The internet is in fact littered with altruistic projects, many of which have now reached the status of essential services, such as Wikipedia, or Github,⁹⁶ (a service for software developers and software version control "where over 100 million developers shape the future of software, together").⁹⁷ Github itself is about to launch its own generative AI product, Copilot - X,⁹⁸ for assisting software developers to solve their problems.

⁹⁴ A comprehensive history of the evolution and meaning of Effective Altruism can be found here. Lewis-Kraus, Gideon. 'The Reluctant Prophet of Effective Altruism'. The New Yorker, 8 August 2022. <https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism>.

⁹⁵ I have pasted the last paragraph of the "History" section of the Wikipedia page documenting the Effective Altruism movement here, so to protect this version from the danger of alteration by a Wikipedia edit rendering a link meaningless for the purpose of this description.

"Following FTX's bankruptcy, the movement underwent additional public scrutiny. Critiques arose not only in relation to Sam Bankman-Fried's role and his close association with William MacAskill, but also concerning issues of exclusion and sexual harassment.[30][39][40][41] In a 2023 Time magazine article, seven women reported misconduct and controversy in the effective altruism movement. They accused men within the movement, typically in the Bay Area, of using their power to groom younger women for polyamorous sexual relationships.[39] The accusers argued that the majority male demographic and the polyamorous subculture combine to create an environment where sexual misconduct can be tolerated, excused or rationalised away.[39] The article also quotes a community liaison from the Centre for Effective Altruism who said that while perpetrators of some of the reported cases have already been banned, the allegations of which they were previously unaware will now be addressed.[39][tone] They also noted that it is challenging to discern to what extent sexual misconduct issues were specific to the effective altruism community or reflective of broader societal problems". https://en.wikipedia.org/wiki/Effective_altruism

⁹⁶ Github provide a free personal use account which allows users to work with unlimited collaborators on unlimited public repositories with a full feature set in perpetuity and provides a similar account for organisations. It also provides paid services.

⁹⁷ <https://github.com/>

⁹⁸ <https://github.com/features/preview/copilot-x>

Much could be learned through examining how to incubate altruistic AI enterprises and much could be done to regulate the current big players to look more like them.

Dr Julia Thornton 26/07/2023