Did the AI train derail?

Carlos A. Afonso -- July 2023

> "Facts are subversive. Subversive of the claims made by democratically elected leaders as well as dictators, by biographers and autobiographers, spies and heroes, torturers and post-modernists. Subversive of lies, half-truths, myths; of all those 'easy speeches that comfort cruel men'."
>
> -- Timothy Garton Ash, *Facts are Subversive*

Recently, hundreds of scientists and people involved in the "algorithm industry" signed a one-sentence manifesto: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[1] A straight-to-the-point warning, even signed by the creators of the fuse that triggered the "generative artificial intelligence" (GAI) crisis: chatGPT. Variants of this alert have been published by institutions and specialist entities, having in common the mantra of "ethical AI".

In an opposite movement, the *Washington Post* reported that Silicon Valley was experiencing a gloomy environment, with mass layoffs, until it was baffled by the GAI tsunami. In the month of May alone, venture investments in AI "start-ups" totaled US$11 billion, an 86% jump compared to the same month last year.[2] This fever, combined with the unbridled prospecting of cryptocurrencies, pushed the main manufacturer of high-performance graphics processors, Nvidia, to the pedestal of multibillion-dollar companies. Graphics processors (GPUs) have been used for fast processing of huge volumes of data, as they have a much higher performance than general-purpose processors – a capacity required by current GAI systems.

It is a curious contradiction between the fear of an AI-caused extinction and the uncontainable desire to make fame and money from its spectacular and frightening advances.

GAI is a variant of a field of systems programming known as Natural Language Processing (NLP), which includes systems such as text generators, customer service chatbots , media conversion and manipulation applications, emulation of biological systems etc.

It is relevant to understand that the origins of AI (whose starting point as a formal research object dates back to 1956)[3] are in computer programming itself, in particular programs that interact with a human user or with another program. Users of devices connected to the Internet (or even offline) interact with a finite state machine at all times and online (or offline) game players interact with variants of fuzzy state machines. You visit a website and look for

[1] See *https://www.safe.ai/statement-on-ai-risk*

[2] See *https://www.washingtonpost.com/technology/2023/06/04/ai-bubble-tech-industry-outlook/*

[3] See *https://en.wikipedia.org/wiki/Dartmouth_workshop*

something of interest in a menu – which represents a finite state machine, with some predetermined options, and you can choose only one. A more sophisticated version (“fuzzy state” or diffuse state) is found, for example, in the interaction in games or with autonomous vehicles, in which the options are dynamic.

These state machines are the forerunners of what is conventionally called artificial intelligence. They were and are nothing more than algorithms in software created by humans. This simplistic explanation is presented just to remind the reader that the fundamentals of AI are in the very genesis of computer programming.

The evolution of processing capacity/speed and memory, as well as the advancement of networked systems, allowed for great leaps in the possibility of increasingly sophisticated interactive programs being able to quickly query large databases distributed in one or more datacenters. This evolution also allowed for large processing and memory capacities to be embedded in a portable computer, a “tablet” or a cell phone, or even in dedicated computers in small devices such as cameras and sensors.

Will Douglas Heaven gave a quick history of the evolution of GAI, showing that the fundamentals came from various teams of developers.[4] One of these fundamentals is the advance, from the 1980s, on the emulation by software of the way in which the neurons of animals interact, forming a neural network, with the capacity to retain and combine information to generate information from textual bases – these are the language models. This advance benefited from an invention by researchers at Google that allowed the meaningful combination of sentences – the "transformers", which enabled recurrent neural networks.

One of the products of these advances was the natural language processor (NLP) "Generative Pre-trained Transformer" (GPT), created in 2018 by the company OpenAI, and which evolved into versions GPT-2, GPT-3 (2020) and GPT-4 or ChatGPT (2022). Its data source is the Internet, bringing to its results all the risks of the quality of information (or misinformation) on the network.

OpenAI's initiative was not the only one. Other software groups besides Google with LaMDA and Bard, Microsoft with a new Bing (using a variant of ChatGPT), as well as a derivative of GPT-3 developed by a consortium of volunteers known as BLOOM, continue to forge ahead in the field of GAI. Meta also produced a variant of the GPT-3 called OPT.

The issues and challenges brought about by these systems provoke an interesting side effect: the emergence of several initiatives that produce legitimizers or detectors of the contents generated by these systems. Chomsky warns that texts resulting from NLPs may be useful for specific niches, but they

[4] Heaven, WD, “ChatGPT is everywhere. Here's where it came from”, *MIT Technology Review* , February 2023.

differ profoundly from how humans reason and use language.[5]

Based on these differences, detectors are being developed, ironically using the same algorithms and information sources, and a review of six existing ones was presented by Funmi Looi Somoye.[6] Some of them, like GPTZero, are even free to use, and claim an accuracy of 96% or better in detecting GAI-generated content. Once this capability is confirmed, the detectors become elements to consider in strategies to combat misinformation or the misuse of content derived from the use of GAI – a challenge especially for the academic environment that seeks to combat plagiarism.

Can we imagine that these possibilities of resistance might mitigate the "end of the world" advocated by the experts who signed the somber one-sentence manifesto mentioned at the beginning of this text? Karen Hao summarizes the nature of the challenges, and note that her text is from May 2021, before the "tsunami" of ChatGPT and the like, highlighting that deviations or abnormalities of humanity are reflected in its algorithms:

> "Studies have already shown how racist, sexist, and abusive ideas are embedded in these models. They associate categories like doctors with men and nurses with women; good words with white people and bad ones with Black people. Probe them with the right prompts, and they also begin to encourage things like genocide, self-harm, and child sexual abuse. Because of their size, they have a shockingly high carbon footprint. Because of their fluency, they easily confuse people into thinking a human wrote their outputs, which experts warn could enable the mass production of misinformation."[7]

Hao also recalls that these large-scale systems devour energy in amounts comparable to large cryptocurrency mining systems. [8]More pessimistic is professor Eugenio Bucci, already under the impact of the ChatGPT buzz:

> "[Generative] AI tools are gradually taking over the discursive protocols that have always guided human behavior. Legal jargon is one such protocol. The scientific method is another. The activity of physicians is a third type. Religions also have their own, which cannot be confused with the previous ones. All of these protocols have one common trait: they are built into the language. When AI learns to speak, as if it were people, it appropriates the protocols that shape individual and social behavior and, from then on, everything changes. As a result, the human being will lose relevance, while dehumanized protocols will expand. From our irrelevance will sprout the vicious cycle that will corner us and then extinguish us. Unless democracy takes action. According to the select group that signed the single-sentence manifesto, there is still time."[9]

[5] Chomsky, N. et al., "The False Promise of ChatGPT", *New York Times* , 08-03-2023.

[6] Somoye, F.L., " ChatGPT detectors in 2023", *PCGuide* , April 2023.

[7] Hao, K., "The race to understand the exhilarating, dangerous world of language AI", *Technology Review* , 20-05-2021. *https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/*

[8] See, for example, Strubell, E., Ganesh, A., McCallum, A., "Energy and Policy Considerations for Deep Learning in NLP", College of Information and Computer Sciences, University of Massachusetts Amherst, 05-06-2019.

[9] Bucci, E., "The Most Intelligent End of the World", *O Estado de São Paulo* , 06-01-2023.

The so-called "social platforms" represented in current regulatory proposals by information searching and message exchanging services, prioritizing larger-scale services such as those offered by companies like Alphabet, Amazon, Meta, Apple, Microsoft, are a part of the biggest challenge -- the range of new services such as those offered by GAI variants , the profusion of applications involving large volumes of financial resources from online casinos (most of them headquartered in tax havens), the challenges for security and privacy in the countless variants of cloud services etc.

There is no scope in current regulatory proposals to encompass the broadness of these new challenges. There is yet another space that these proposals are far from reaching: the increasingly diverse universe of the Internet of Things (IoT). In this space there is an infinite variety of devices whose origin is not clear, where the responsibility for the embedded software ("firmware") is difficult to determine, and where security risks are consequently not mitigated by the manufacturers.

Firmware updates on billions of IoT devices are almost non-existent. Nor is there any clarity about the functionality of these "firmwares" – to whom a Wi-Fi camera actually sends the captured images, what kind of non-perceptible interaction a homemade digital assistant has with its manufacturer and so on.

In short, there is a great risk that regulatory proposals, if enshrined in law, will already be born obsolete, or reach a smaller part of the Internet's interactive space. In particular, a serious challenge now appears with GAI. If there were doubts about the worrying impact on copyright and labor rights of this new mode of interaction involving gigantic databases captured (legally or illegally) from the Internet as well as sophisticated software, the current example of the strike movement in Hollywood eliminates them.[10]

Actors and actresses have their interpretations usurped by companies that reproduce their original performances digitally in other performances, often without authorization from these artists. Screenwriters have their original texts usurped by the use of GAI to generate new writings by the studios, without remunerating the human authors.

These are risks whose consequences will still be more precisely evaluated, when the dust from the current explosion of these new modalities of interaction with GAI settles.

[10] *https://en.wikipedia.org/wiki/List_of_Hollywood_strikes*