2023

Will AI take us into Orwell's 1984?

How generative AI could lead to normalization of thinking

AUTHOR
Albert Meige
Director, Blue Shift

Inspired by an interview with Alix Boulnois, Chief Digital Officer and executive committee member of the Accor group

BULLETIN
BLUE SHIFT
BY ARTHUR D. LITTLE

"Generative artificial intelligences (AIs) are based on neural networks that have, on the one hand, been trained on immense bodies of data and information, and on the other hand, been adjusted by humans to return the expected answers. The results are staggering and the performance gains unprecedented, but then so are the risks. In this Blue Shift Bulletin, inspired by an interview with Accor Chief Digital Officer (CDO) and executive committee member Alix Boulnois, we consider briefly one of the darker prospects for AI.

"Orwell, *1984*," was Alix's response when I asked her what she thought would be the most dystopian or least desirable scenario with the advent of AI, particularly generative AI. Alix is the CDO of global hotel group, Accor, and is an expert in major transformations in a technological and digital context. This came up during a conversation we were having with Alix on the potential of AI for the hospitality sector (covered in a separate Blue Shift Bulletin).

Is normalization of thinking really a risk?

For Alix, one of the major risks of generative AI is "normalization of thinking." As we explored in our article "My kids have replaced me with ChatGPT," a primary limitation of generative AIs (as with AIs in general) is algorithmic bias. Algorithmic bias refers to the differences in treatment an AI will make, in a systematic and unintentional way, between individuals or groups of individuals. These biases reflect human biases and are inherent in deep learning and the way neural networks are trained. In that article, we illustrated this fact with a rather disturbing example in which ChatGPT suggested that certain people should be imprisoned based on their country of origin. Safeguards have of course been put in place to limit such algorithmic biases, but these safeguards are also themselves defined by humans. The nature of the AI is therefore a reflection of those who train it.

Of course, safeguards are a very necessary aspect of any AI. Many users have already tested how well ChatGPT and other AIs respond to requests with malicious intent, such as how to devise a terrorist attack or build a weapon, and with some exceptions the safeguards have been shown to be just about OK — so far. (And by "just about OK," I mean that they generally work but have been cracked a few times; it's an ongoing process.)

Related to safeguards is the issue of online hate and the Internet's contribution to the increasing polarization of society, as reflected in many social media exchanges. AI can be a powerful tool to help combat online hate, with its ability to quickly identify where it occurs and instantly take action, such as formulating a suitable response or quarantining or removing content. This can greatly improve the efficiency and effectiveness of human mediators who are faced with oceans of new input every day.

But some worry about the risk of normalization of thinking that these safeguards imply. For example, who decides what is and is not acceptable? In the case of ChatGPT and other AIs, decisions about acceptability are effectively made behind closed doors by the employees of a large tech corporation. Elon Musk, cofounder of OpenAI (the company behind ChatGPT), expressed his own concerns on Twitter in February 2023:

> OpenAI was created as an open source (which is why I named it *Open*AI), nonprofit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft. Not what I intended at all.

As Asma Mhalla, lecturer at SciencesPo Paris and École Polytechnique, recently said, "Technology is the vehicle of normalization." This has been the case since the first p rinting presses made mass production of books possible, and generative AI is no exception. Its designers shape the tools according to their own worldview — a historical, economic, strategic, political, ethical, and/or philosophical vision of the world.

"OpenAI has become a closed source, maximum profit company ... Not what I intended at all."

Elon Musk, cofounder, OpenAI

For a developer whose vision is libertarian, for example, their AI likely will be as well. It might allow freedom of expression that employs speech and delivers content that some would find unacceptable. Likewise, with a more authoritarian vision, a developer's AI may follow suit, banning the use of certain words and content that could offend perhaps even very small minorities.

Why is this a major risk? When you use a classic search engine such as Google, the engine returns huge numbers of documents. Even if the ranking algorithm is biased, which may be hard to avoid in practice, with some effort it is possible to read the pros and cons on a given subject and form your own opinion. However, an AI like ChatGPT produces *unsourced synthesis*. This synthesis inextricably and invisibly embeds its designer's world vision. The information is predigested according to that vision. It saves time, of course, but it is time we otherwise would have used for reflection. So, we save time, but we lose reflection.

Global reactions — such as the decision by Italy to ban the use of ChatGPT due to privacy concerns as well as the open letter signed by more than 1,000 AI experts and executives calling for a pause in further system development — reveal the extent of the concern. The question raised is fundamental. It is a question about who owns our collective vision of the world and who has sovereignty over our collective values.

As individual ChatGPT users, most of us want a tool that is safe and secure. Most of us also value an online environment that does not exacerbate hate and avoids driving further polarization of attitudes. But as AI increasingly permeates everyday interactions, at what point do the implicit values and attitudes embedded into AI synthesis begin to stifle free thought? Will ubiquitous AI start to hinder the healthy confrontation of ideas, a basic prerequisite for progress? These questions need to be considered. Unfortunately, they will be explored a *posteriori*, now that these generative AIs are already spreading like wildfire in our companies and even in our homes. We are, collectively, presented with a fait accompli.

> **"Most of us value an online environment that does not exacerbate hate and avoids driving further polarization of attiudes."**

Conclusion
Back to the future

For as long as I can remember, I've been fascinated by technology. At the same time, I've always had some apprehension about it. I remember reading Aldous Huxley's *Brave New World* as a teenager. In this classic dystopian science fiction novel, Huxley presents a society that, through technology, has become very stable, without wars or quarrels. But this apparent harmony is unfortunately built at the expense of individual freedom and free will. And it's true that technology's impact is often ambivalent; social media is a good example of this.

This ambivalence could not be more worrying when it comes to AI. As Alix mentioned, the people in George Orwell's novel *1984* are conditioned by a totalitarian state, with the Ministry of Truth tailoring reality and history. Its "Newspeak" output is a language without nuance, a language that shapes thought — and consequently leads to the absence of thought.

Regulation always trails new technology development and is usually established in response to new incidents or crises that result from its widespread adoption. It is not hard to envisage a tightening of controls that could ensue from AI-related incidents in the coming months and years. Will AIs like ChatGPT inevitably lead us toward Newspeak? Does ChatGPT signal the end of critical thinking and reflection? And if that's our future, what are we going to do about it?

Blue Shift, by Arthur D. Little, explores the impact of technologies on business, society, and humans.

For further information, visit: www.adlittle.com

BLUE SHIFT

BY ARTHUR D. LITTLE