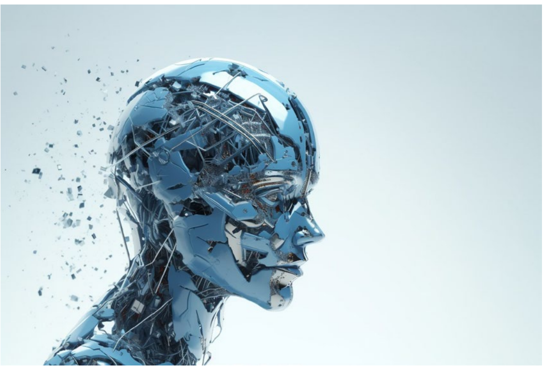
What if generative artificial intelligence became conscious?

Generative artificial intelligence applications, such as ChatGPT, are powered through complex learning processes by comprehensive datasets of – potentially dubious – human-created content. There are concerns that such tools could develop consciousness and spark emergent behaviour that is, by definition, unpredictable and therefore potentially unsafe. Do these concerns point to a need to look again at the relevant legislation?

Artificial intelligence (AI) is a very generic and far-reaching concept that is often used when discussing more specific technologies, such as machine learning, neural networks and deep learning. Recent discussions around the possibility of a somehow unexpected or 'emergent' behaviour associated with AI, really have to do with deep learning in neural networks, since these refer to generic structures that do not incorporate set rules or decision trees (or other tools such as genetic algorithms) but need to be trained on coherent data sets to eventually produce proper results.



© Diana Vyshniakova / Adobe Stock

Neural networks are structures inspired by the human brain. They are composed of a colony of elementary cells, or neurons, that are massively interconnected. Through various training methods, the information providing the desired behaviour is stored as a set of 'weights' in those interconnections. 'Weight' here refers to the strength of the signal at the receiving neuron when the interconnection is activated by the neuron at the other end of that specific link. Neurons are simple cells whose main function is to pass on the signals according to weights. They also have to adjust those weights, in an ongoing learning process, according to a system of rewards depending on the quality of the overall results, which are, to some extent, unpredictable. Consciousness can eventually spark at a higher level in colonies of cells that are big enough, as is the case for the human brain.

Ant colonies and beehives have a similar model of interconnected nodes without individuality. In these cases, however, the level of consciousness seems very limited, perhaps because the interactions between individuals are very limited too, despite showing characteristics of emergent behaviour. In this sense, a network of humans corresponds much more closely to the conditions found in a neural network because, thanks to language, the links are stronger and the communication much better. Although networks of people do not seem to have consciousness, human society – and its overall behaviour – can become very complex, and has empowered humankind with decisive tools, such as civilisation, through a very deep sharing of knowledge. This is behind the idea of 'collective consciousness' expressed by Durkheim: 'The totality of beliefs and sentiments common to the average members of a society forms a determinate system with a life of its own. It can be termed the collective or common consciousness'. Collective consciousness is not unique and can vary, for instance, between different nationalities.

The real question may be how to define consciousness. Since Freud divided human consciousness into three levels of awareness: unconsciousness, preconsciousness and consciousness, philosophers have described consciousness and its levels in varying ways, from mere self-awareness and survival instinct to societal contribution. Zukerfield highlighted the concept of self in relation to society, and others, such as Seth, concentrate on individual emotions. Gramsci called 'collective consciousness' a 'living organism' and related it to the class struggle and solidarity movements among oppressed people. Other authors refer to 'collective imagination' to stress the relationship between social movements and policy.

Potential impacts and developments

Generative AI applications (that use models based on neural networks to identify patterns and structures within existing data to generate new and original content) have become very popular, but also controversial, with the recent deployment of new versions of ChatGPT and other new large language models (LLM). While artificial intelligence in general has been the focus of attention and the subject of legislative work at EU level for some time, the release of new generative systems has shown that technological change may be happening too fast for EU law to keep up. However, as was the case with search engines such as Google, some chatbots have appeared as data pre-processing interfaces, meeting the gap in the market caused by the exponential rise in the volume of information available on the internet and our limited capacity, as humans, to process it. Generative AI is here to stay.

Since generative AI is usually trained on data available on the internet, it is bound to mirror the overall collective consciousness of the society generating that data. It is that collective consciousness that is reflected, for instance, in chatbots based on generative AI. However, there is no guarantee that a collective consciousness will always lead to wise decisions. For example, a mob will do things that most of its individual members would never do alone. Although collective consciousness may have no sense of 'collective remorse', the individuals constituting it would blame the mob as an independent entity. Inadequate decisions based on biased information can be even more harmful than malicious or selfish acts; and they are usually harder to assess. The danger is that a conscious AI might have spurious motivations such as gaining more 'likes' on Twitter (now X), potentially fuelling dangerous populism.

For historian Yuval Noah Harari, there is a real danger that sentient machines, having mastered language, will go on to influence people and destroy our ability to have meaningful conversations, leading to the destruction of democracy. However, the question now is not whether a single computer can do this, since that one computer can be controlled or curated in some way, but whether a group of interconnected computers could behave as a neural network and gain a separate uncontrolled consciousness residing in the cloud, leading to emergent behaviour.

Anticipatory policymaking

Neurons are very basic cells without individual consciousness, whereas human society is composed of conscious individuals. While computers do not have individual consciousness, they may incorporate alternative data sets and act according to specific rules interfering with the network at a higher level. This is why a network of computers would behave in a way that is more comparable to a network of individuals, to a society or, considering all the shared previous knowledge and rules of conduct, even to a culture or civilisation. This should leave generative AI far from individual consciousness and shy of Harari's apocalyptic scenarios or examples depicted in film, such as 'Skynet' in *Terminator*, 'Hal' in *2001: A space odyssey*, and the spaceship autopilot in *Wall-e*.

However, even the most advanced societies are not safe from generating 'mobs' (think of the attack on the US Capitol in 2021). Arguably, incomplete and biased information, fake news and disinformation can feed back, producing a magnifying effect through generative AI that may be unpredictable. The regulation of artificial intelligence must be closely linked to proper regulation of data and its governance and the concept of liability. After recently passing legislation such as the Digital Services Act, the Digital Markets Act, the Data Act and the Data Governance Act, the EU is now finalising work on the artificial intelligence act. This will define generic mechanisms to regulate the application and development of AI, assess how this regulation should be applied to generative AI and help decide whether additional regulatory instruments are needed.

The European Parliament will be monitoring implementation of the EU's Horizon programme, which facilitates and funds research on artificial intelligence among other things. It will also oversee the 2030 'Path to the digital decade' policy programme, with the target of 75 % of EU companies using the cloud/AI/big data by 2030. The EU will also continue to shape policies in areas strongly affected by generative AI, such as the creative economy, education, health and many more industrial, social and cultural domains.

What-ifs are two-page-long publications about new or emerging technologies aiming to accurately summarise the scientific state-of-the-art in an accessible and engaging manner. They further consider the impacts such technologies may have – on society, the environment and the economy, among others – and how the European Parliament may react to them. As such, they do not aim to be and cannot be prescriptive, but serve primarily as background material for the Members and staff of the European Parliament, to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

stoa@ep.europa.eu (contact) http://www.europarl.europa.eu/stoa http://www.europarl.europa.eu/thinktank (internet) http://epthinktank.eu (blog)