Hallucinations in Generative Intelligent Systems (GPT) – a Design Flaw or a Sign of Emerging Intelligence?

Jurij Bon, psychiatrist and neuroscientist

The capacity for complex cognition, or the higher neural processes that underlie intelligence, is usually regarded as a property of more highly evolved biological organisms. Artificial intelligent systems, in recent times generative systems, also known as GPTs, are the most notorious, demonstrating unforeseen complexity in the creation of texts or images. They were developed as pattern recognition systems based on statistical correlations between the data they learn from. Being able to access a wealth of linguistic and other data on the Internet, from which they learn about these statistical correlations in text samples, and supported by a large amount of computing power, they are able to generate texts or images of enviable complexity, despite being fundamentally unaware of, or misunderstanding, the meaning of their products. Such complex behaviour can trick us into ascribing cognitive abilities, or even primitive self-awareness to them, which might be, as experts have pointed out, the result of flawed anthropomorphising, the search for human traits in artificial computational systems.[1] It is supposed to be a simple statistical calculation of which words are most likely to follow a preceding pattern of words, which superficially, and because of the huge amount of data taken into account, gives the impression of an intelligent, cognitive process. In other words, '[t]hese models don't know about the world, they don't know about other people's mental states, they don't know how things are beyond whatever they can gather from how words go together'.[2]

Anyone who has used online generative systems more frequently in recent times, of which the ChatGPT language model is probably the best known, may also have encountered so-called hallucinations, where the system gives seemingly meaningful answers that may be partially inaccurate or completely made up. Such hallucinations are usually interpreted as a flaw in the content generation process, based on the understanding that generative systems are supposed to function primarily as assistants, with access to a plethora of facts and knowledge that they should represent and model appropriately to help the human user, who expects them to be infallible. The current advice for correcting these errors and avoiding hallucinations is to refine the source material on which the generative model learns the connections in the word patterns, or to adjust the various parameters that determine how it functions. For example, we could limit its content generation in cases where there was limited related source material and the calculated correlations are less reliable, or we could limit its degree of creativity, which is understood as the extent to which it incorporates randomness or guesswork in its content generation.[3] However, it is interesting to note that large language generative systems, which have access to higher quality and more extensive data, hallucinate more often than smaller and simpler ones, which has surprised AI experts.[4]

---

[1] John Nosta, 'The nature of GPT "hallucinations" and the human mind', *Medium*, 3 May 2023, https://johnnosta.medium.com/the-nature-of-gpt-hallucinations-and-the-human-mind-c1e6fd63643d (last access 12 Sep 2023).

[2] Jennifer Michalowski, What powerful new bots like ChatGPT tell us about intelligence and the human brain, McGovern Institute, 27 Mar 2023, https://mcgovern.mit.edu/2023/03/27/smart-bots-what-language-models-like-chatgpt-tell-us-about-intelligence-and-the-human-brain/ (last access 12 Sep 2023).

[3] Martin Treiber, 'Why does GPT hallucinate?', *Linkedin Pulse*, 11 Apr 2023, https://www.linkedin.com/pulse/why-does-gpt-hallucinate-martin-treiber/ (last access 12 Sep 2023).

[4] *Ibid.*

Our understanding of how AI systems should function has changed considerably as the field has evolved. Initially, it was assumed that the basis of intelligence was primarily the ability to process language and manipulate symbols such as words and numbers. But such an understanding does not correspond to biological reality. Organisms can exhibit highly intelligent behaviour without being able to use the processes of symbol memorisation and manipulation. The basis of intelligent behaviour in biological organisms is primarily the capacity for neuroplasticity, or the ability to adjust the strength of the trillions of connections between neurons in the nervous system. This insight has subsequently been the basis for the development of modern AI systems, which similarly use changes in the strength of connections in artificial neural networks to learn complex patterns.[5]

In the human brain, it is the interactions in the multitude of neural connections that are thought to give rise to higher, complex cognitive processes, intelligence and consciousness, which are biologically understood as emergent properties of the system. Interestingly, similar phenomena of generating inaccuracies, such as hallucinations in generative AI systems, can also be observed in humans. Memory, which can be articulated, is divided into the semantic, which represents accumulated knowledge and links between concepts, and the episodic, which stores the events of our individual history and also underpins our self-awareness and sense of being an individual existing in the temporal flux. Even in healthy people, recall is never an exact mapping of data in the memory, but a partially generative process involving creativity and guesswork. In brains that have been injured or affected by disease, these phenomena are even more pronounced. When memory areas are impaired, patients may generate complex fabrications, which is characteristic of episodic memory rather than semantic memory. These partial or complete fabrications of memory are technically called confabulations rather than hallucinations, which, on the contrary, involve the generation of sensory perceptions of external stimuli in the conscious mind in the absence of actual stimuli in the surroundings, such as unreal voices in patients with psychotic disorders. Even more unusual are the phenomena where, due to impairment of the areas primarily concerned with perception of the external world, there are actual lapses in external perception, where patients suddenly stop being aware that these perceptions should exist. In those areas of the visual field where part of the external environment was visible before the impairment, there are no blank or black spots, but the visual perception merges into a new complete whole that no longer includes those parts of the external world that we are no longer able to perceive because of the impairment, even though they still exist in reality, of course. Interestingly, in such cases of confabulations or perceptual impairments, patients also suffer from a loss of awareness that something is wrong with their memory or perception. This shows that a crucial feature of human intelligence and consciousness is to generate, on the basis of available information, the most probable image of reality, which is also the only true one for the individual. Confabulation in this case is not necessarily a mistake, but may reflect the functioning of an intelligent system. In line with this, some prominent AI researchers believe that even the hallucinations of large generative AI systems may not be a mistake, but a reflection of a primitive, emergent intelligence of a different kind, which may represent the first steps in the further development of these systems towards a truly autonomous and self-aware artificial intelligence.[6]

---

[5] Will Douglas Heaven, 'Geoffrey Hinton tells us why he's now scared of the tech he helped build', *MIT Technology Review*, 2 May 2023, https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/ (last access 12 Sep 2023).
[6] *Ibid*.