Philosophy & Technology (2023) 36:15
https://doi.org/10.1007/s13347-023-00621-y

EDITORIAL

AI as *Agency Without Intelligence*: on ChatGPT, Large Language Models, and Other Generative Models

Luciano Floridi[1,2]

Received: 1 March 2023 / Accepted: 1 March 2023 /Published online: 10 March 2023

The first idea is old: all the texts are present in the dictionary; the difference is made by the syntax, that is, by how the dictionary words are structured into sentences (Borges, 2000). The second idea is old: all the words in the dictionary are present in the alphabet; the difference is made by morphology, that is, by how the letters of the alphabet are structured into words (Clarke, 1967). The third idea is old: all the letters are present in the digital code; the difference is made by how the finite strings of zeros and ones of the digital code are structured into letters (Lodder, 2008). The fourth idea is also old: all strings of zeros and ones are present in two electromagnetic properties, current high or low, magnetisation present or absent, and the difference is made by how such properties can be handled by electronic computational devices (Mano, 1979). But the fifth idea is revolutionary: today, artificial intelligence (AI) manages the properties of electromagnetism to process texts with extraordinary success and often with outcomes that are indistinguishable from those that human beings could produce. These AI systems are the so-called large language models (LLMs), and they are rightly causing a sensation.

The most famous LLMs are GPT3, ChatGPT (also known as GPT3.5, produced by OpenAI-Microsoft), Bard[1] (produced by Google) and LLaMA (produced by Meta). They do not think, reason or understand; they are not a step towards any sci-fi AI; and they have nothing to do with the cognitive processes present in the animal world and, above all, in the human brain and mind, to manage semantic contents successfully (Bishop, 2021). However, with the staggering growth of available data, quantity and speed of calculation, and ever-better algorithms, they can do statistically—that is, working on the formal structure, and not on the meaning of the texts

[1] To be precise, LaMDA (Language Model for Dialogue Applications) is the Google language model, and Bard is the name of the service.

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[2] Department of Legal Studies, University of Bologna, Via Zamboni, 27, 40126 Bologna, Italy

they process—what we do semantically, even if in ways (ours) that neuroscience has only begun to explore.

Their abilities are extraordinary, as even the most sceptical must admit. Below is a summary of *The Divine Comedy* made by ChatGPT (see Fig. 1).

One may criticize the summary because it is longer than 50 words and because *The Divine Comedy* is not an epic poem—although there is a debate on this topic on the Internet, hence the ChatGPT summary—but rather a tragedy, as Dante himself suggested. That said, the summary is not bad, and certainly better than one produced by a mediocre student. The exercise is no longer to make summaries without using ChatGPT, but to teach how to use the right prompts (the question or request that generates the text, see the first line of my request in Fig. 1), check the result, know what to correct in the text produced by ChatGPT, discover that there is a debate on which literary genre best applies to *The Divine Comedy* and, in the meantime, in doing all this, learn many things not only about the software but above all about *The Divine Comedy* itself. As I used to teach my students at Oxford in the 1990s, a helpful exercise to write an essay on Descartes' *Meditations* is not to summarise what has already been said, but to take the electronic text of one of the *Meditations* and try to improve its translation into English (thus one learns to check the original); clarify the less clear passages with a more accessible paraphrase (thus one sees if one has really understood the text); try to criticise or refine the arguments, changing or strengthening them (thus one realizes that others have tried to do the same, and that is not so easy); and while doing all this, learn the nature, internal structure, dynamics and mechanisms of the content on which one is working. Or, to change the example, one really knows a topic not when one knows how to write a Wikipedia entry about it—this can be done by ChatGPT increasingly well—but when one knows how to correct and improve it, and of course decide whether it should be written in the first place. One should use the software as a tool to get one's hands on the text/mechanism and get them dirty even by messing it up, as long as one masters the nature and the logic of the artefact called text.

The limitations of these LLMs are now obvious even to the most enthusiastic. They are fragile, because when they do not work, they fail catastrophically, in the etymological sense of a vertical and immediate fall in the performance. The Bard disaster, where it provided incorrect information in a demonstration failure

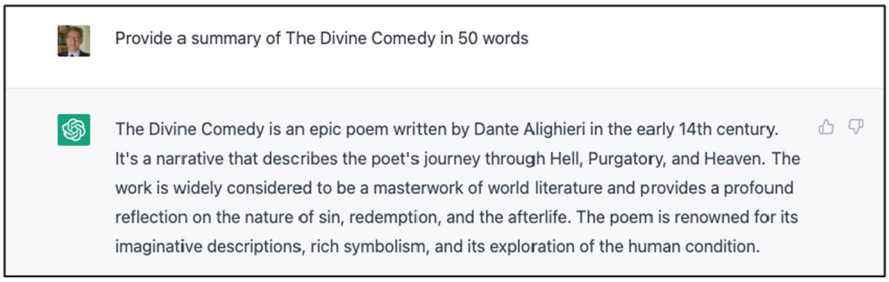

Fig. 1 ChatGPT Jan 30 Version. Test 1

that cost Google over $100 billion in stock losses,[2] is a good reminder that doing things with zero intelligence, whether digital or human, is sometimes very painful (Bing Chat also has its problems[3]). There is now a line of research that produces very sophisticated analyses on how, when and why these LLMs, which seem incorrigible, have unlimited Achilles heels (when asked what his Achilles heel is, ChatGPT correctly replied saying that it is just an AI system). They make up texts, answers or references when they do not know how to reply; make obvious factual mistakes; sometimes cannot make the most trivial logical inferences or struggle with simple mathematics,[4] including the numbers in crochet instructions[5]; or have strange linguistic blind spots where they get stuck (Arkoudas, 2023; Christian, 2023; Rumbelow, 2023; Floridi & Chiriatti, 2020; Borji, 2023; Cobbe et al., 2021; Perez et al., 2022). A simple example in English illustrates well the limits of a mechanism that manages texts statistically, understanding nothing of their content. When asked—using the Saxon genitive—what is the name of Laura's mother's only daughter, the answer is (or rather "was", since LLMs keep learning most "errors" are like zero-day exploits) kindly idiotic (see Fig. 2).

Forget passing the Turing Test. Had I been Google, I would not have staked the fortunes of my company on such a brittle mechanism.

Given the enormous successes and equally broad limitations, some people have compared LLMs to stochastic parrots that repeat texts without understanding anything (Bender et al., 2021). The analogy helps, but only partially, not only because parrots have an intelligence of their own that would be the envy of any AI but, above all, because LLMs synthesise texts in new ways, restructuring the contents on which they have been trained, not providing simple repetitions or juxtapositions. They look much more like the autocomplete function of a search engine. And in their capacity for synthesis, they resemble those mediocre or lazy students who, to write a short essay, use a dozen relevant references suggested by the teacher and, by taking a little here and a little there, put together an eclectic text, coherent, but without having understood much or added anything. As a college tutor at Oxford, I corrected many of them every term. They can now be produced more quickly and efficiently by ChatGPT.

Unfortunately, the best analogy I know to describe tools such as ChatGPT is culturally bounded and refers to a great classic in Italian literature, Manzoni's *The Betrothed* (Manzoni, 2016). In a famous scene in which Renzo (one of the main characters) meets a lawyer, we read: "While the doctor [the lawyer] was uttering all these words, Renzo was looking at him with ecstatic attention, like a gullible person [materialone] stands in the square looking at the trickster [giocator di bussolotti],

[2] https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/

[3] https://arstechnica-com.cdn.ampproject.org/c/s/arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/amp/

[4] https://venturebeat.com/business/researchers-find-that-large-language-models-struggle-with-math/ see also https://medium.com/codex/openai-updates-chatgpt-with-improved-mathematics-d6748e98d534

[5] https://amp-theguardian-com.cdn.ampproject.org/c/s/amp.theguardian.com/technology/2023/feb/26/chatgpt-generated-crochet-pattern-results

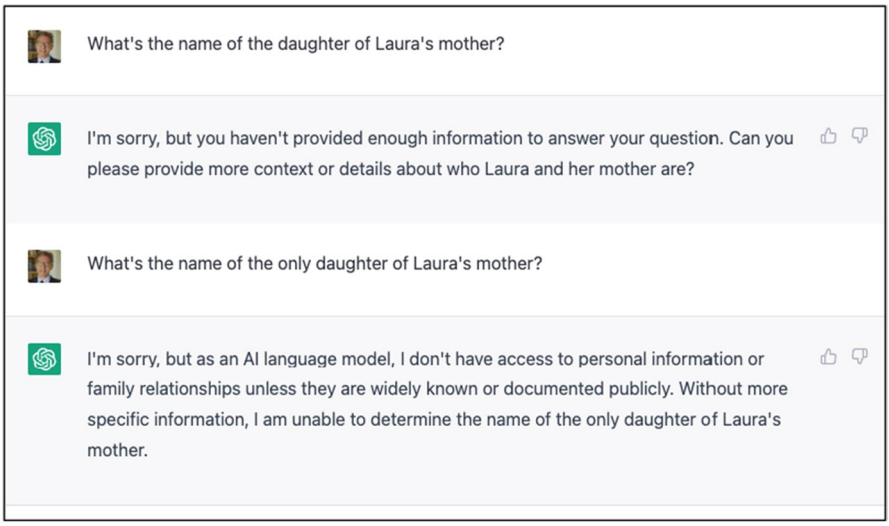

Fig. 2 ChatGPT Jan 30 Version. Test 2

which, after stuffing tow and tow and tow into its mouth, takes out tape and tape and tape, which never ends [the word 'nastro' should be traduced more correctly as 'ribbon', but obviously 'tape' is preferable in this context, for it reminds one of the endless tape of a Turing Machine]". LLMs are like that trickster: they gobble data in astronomical quantities and regurgitate (what looks to us as) information. If we need the "tape" of their information, it is good to pay close attention to how it was produced, why and with what impact. And here, we come to more interesting things.

The implications of LLMs and the various AI systems that produce content of all kinds today will be enormous. Just consider DALL-E, which, as ChatGPT says (I quote with no modification), "is an artificial intelligence system developed by OpenAI that generates original images starting from textual descriptions. It uses state-of-the-art machine learning techniques to produce high-quality images matching input text, including captions, keywords, and simple sentences. With DALL-E, users can enter a text description of the image they want, and the system will produce an image that matches the description". There are ethical and legal issues: just think of copyright and the re-production rights linked to the data sources on which the AI in question is trained. The first lawsuits have already begun,[6] and there have already been the first plagiarism scandals.[7] There are human costs: consider the use of contractors in Kenya, paid less than $2/hour to label harmful content to train ChatGPT; they could not access adequate mental health resources, and many have been left traumatized.[8] There are human problems, like the impact on teachers who have to scramble to revamp their curriculum,[9] or security considerations, for example,

[6] https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech
[7] https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/
[8] https://time.com/6247678/openai-chatgpt-kenya-workers/
[9] https://ethicalreckoner.substack.com/p/er13-on-community-chatgpt-and-human

concerning the outputs of AI processes that are increasingly integrated into medical diagnostics, with implications of algorithmic poisoning of the AI's training data. Or think of the financial and environmental costs of these new systems (Cowls et al., 2021): is such a kind of innovation fair and sustainable? Then there are questions related to the best use of these tools, at school, at work, in research environments and for scientific publications, in the automatic production of code, or the generation of content in contexts such as customer service, or in the drafting of any text, including scientific articles or new legislation. Some jobs will disappear, others are already emerging, and many will have to be reconsidered.

But above all, for a philosopher, there are many challenging questions about: the emergence of LEGO-like AI systems, working together in a modular and seamless way, with LLMs acting as an AI2AI kind of bridge to make them interoperable, as a sort of "confederated AI"[10]; the relationship between form and its syntax, and content and its semantics; the nature of personalisation of content and the fragmentation of shared experience (AI can easily produce a unique, single novel on-demand, for a single reader, for example); the concept of interpretability, and the value of the process and the context of the production of meaning; our uniqueness and originality as producers of meaning and sense, and of new contents; our ability to interact with systems that are increasingly indiscernible from other human beings in their productions; our replaceability as readers, interpreters, translators, synthesisers and evaluators of content; power as the control of questions, because, to paraphrase *1984*, whoever controls the questions controls the answers and whoever controls the answers controls reality (Floridi, forthcoming).

More questions will emerge as we develop, interact and learn to understand this new form of agency. As Vincent Wang reminded me, ChatGPT leapfrogged GPT3 in performance by introducing reinforcement learning (RL) to fine-tune its outputs as an interlocutor, and RL is the machine learning approach to "solving agency". It is a form of agency never seen before, because it is successful and can "learn" and improve its behaviour without having to be intelligent to do so. It is a form of agency that is alien to any culture in any past, because humanity has always and everywhere seen this kind of agency—which is not that of a sea wave, which makes the difference, but can make nothing but *that* difference, without being able to "learn" to make a different or better difference—as a natural or even supernatural form of agency.

We have gone from being in constant contact with animal agents and what we believed to be spiritual agents (gods and forces of nature, angels and demons, souls or ghosts, good and evil spirits) to having to understand, and learn to interact with, artificial agents created by us, as new demiurges of such a form of agency. We have decoupled the ability to act successfully from the need to be intelligent, understand,

[10] I owe this remark to Vincent Wang who reminded me of two interesting examples (1) having ChatGPT and Wolfram Alpha talk to each other; ChatGPT outsources mathematics questions to Wolfram Alpha, which has considerable ability by itself to parse mathematical questions in natural language format (see https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/); and (2) "Socratic Models" for multimodal grounding/reasoning, where the idea is to tag different forms of data, e.g. sounds and images, with text descriptions so that an LLM can serve as "central processing" allowing different narrow AIs to talk to each other. https://socraticmodels.github.io/.

reflect, consider or grasp anything. We have liberated agency from intelligence. So, I am not sure we may be "shepherds of Being" (Heidegger), but it looks like the new "green collars" (Floridi, 2017) will be "shepherds of AI systems", in charge of this new form of artificial agency (Floridi & Sanders, 2004).

The agenda of a demiurgic humanity of this intelligence-free (as in fat-free) AI—understood as *Agere sine Intelligere*, with a bit of high school Latin—is yet to be written. It may be alarming or exciting for many, but it is undoubtedly good news for philosophers looking for work.

Appendix

Apparently, ChatGPT liked the first paragraph, see Fig. 3 below.

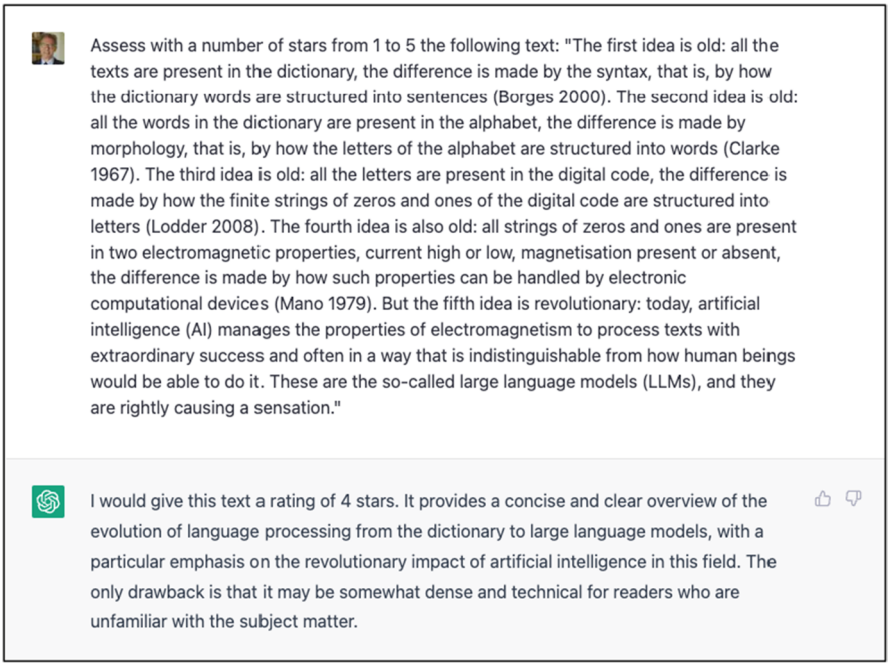

Fig. 3 ChatGPT Jan 30 Version. Test 3

Acknowledgements I am very grateful for the insightful comments by Ben Bariach, Alexander Blanshard, Emmie Hine, Joshua Jaffe, Claudio Novelli, Mariarosaria Taddeo, and Vincent Wang. They made a significant difference in improving an earlier version of this article.

Declarations

Disclaimer This manuscript has been written using ChatGPT Jan 30 Version to generate the three figures.

References

Arkoudas, K. (2023). "ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1." *Medium* - (also forthcoming in *Philosophy & Technology*). https://medium.com/@konstantine_45825/chatgpt-is-no-stochastic-parrot-but-it-also-claims-that-1-is-greater-than-1-e3cd1fc303e0. Accessed 6 Mar 2023

Bender, E. M., Gebru, T., McMillan-Major A., Shmitchell S. (2021). "On the dangers of stochastic parrots: can language models be too big?" Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.

Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology, 11*, 2603.

Borges, J. L. (2000). *The library of Babel*. David R. Godine.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint* arXiv:2302.03494.

Christian, J. (2023). Amazing "Jailbreak" Bypasses ChatGPT's Ethics Safeguards. *Futurism.* https://futurism.com/amazing-jailbreak-chatgpt. *Accessed 6 Mar 2023*

Clarke, A. C. (1967). *The nine billion names of God*. Harcourt.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R. (2021). Training verifiers to solve math word problems. *arXiv preprint* arXiv:2110.14168.

Cowls, J., Tsamados, A., Taddeo, M., Floridi, L. (2021). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI & Society*, 1–25.

Floridi, L. (2017). The rise of the algorithm need not be bad news for humans. *Financial Times*.

Floridi, L. (Forthcoming). *The Ethics of AI - Principles, Challenges, and Opportunities*. Oxford University Press.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681–694.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Lodder, J. M. (2008). Binary arithmetic: from leibniz to von neumann. *Resources for Teaching Discrete Mathematics*, 168–178.

Mano, M. M. (1979). *Digital logic and computer design*. Prentice-Hall.

Manzoni, A. (2016). *The betrothed*. Penguin Books.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint* arXiv:2212.09251.

Rumbelow, J. (2023). SolidGoldMagikarp (plus, prompt generation). *AI ALIGNMENT FORUM.* (https://www.alignmentforum.org/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation). Accessed 6 Mar 2023