# Ghosts in the AI

Emanuele Fulvio Perri[1, †], Elio Grande[2, †]

[1, 2] *University of Pisa, Largo Bruno Pontecorvo, 3, 56127, Pisa, Italy*

## Abstract

This work regards the social side of trustworthiness in the context of Large Language Models (LLMs) according to two congruent shades. Indeed, the first paragraph, drawing aid from a passage of *The Science of Logic* by G. W. F. Hegel, proposes a qualitative and semantic interpretation of the origin of the so-called "emergent abilities" of LLMs, which are deemed something more complex than a trivial deceit. The second paragraph rather concerns the topic of trustworthiness and responsibility of LLMs from an ethical and phenomenological perspective, proposing a parallelism between the issue of extended mind and the generative transformers as a cognitive extension. The focus lies on the repercussions for the intensive utilization, which can be summarized in the concepts of cognitive depletion and digital dementia, leading to a debasement of precious human qualities – creativity, attention, interpretational ability. Our suggestion, then, first of all trusting—because we *have to* trust—the critical sense of human users, is directed towards some kind of ethics *of* AI to introduce in the K-12 category. Our aim remains the wished for design of a pacific coexistence.

## Keywords

Generative AI, emergent abilities, extended mind, hallucinations, cognitive depletion.

## 1. Introduction

A deviation had occurred at the last mile, in the long run of the approval of the Artificial Intelligence Act, because of an unexpected technological evolution: the so-called Foundation Models, generative artificial intelligence devices made of deep neural networks good enough to elaborate coherent responses to input prompts, concerning many typologies of data and particularly processing natural language within diverse conceptual and linguistic domains. The definitive text of the AI Act – see in particular article 51 and annex XIII – provides some criteria of "systemic risk" for general purpose models, among other things, in the number of parameters of the models, in the quality and dimension of datasets and above all in the necessary compute for training, fixing the plausible risk threshold to 10^25 FLOPs [16]. We will follow some suggestions regarding the origin of the so-called "emergent abilities" of Large Language Models (LLMs)[2], developing them through some considerations about the extensions of the mind. If there is a character which is bearer of risk in LLMs, it is their everyday pervasiveness. From *Una domanda impossibile ad Artemisia Gentileschi* ["*An impossible question to Artemisia Gentileschi*"], the Turing test on a sample of more than 1200 participants distributed by various age and education, jointly conceived in 2023 by the Departments of Computer Science and Civilization and Forms of Knowledge of the University of Pisa, it has emerged that 31,5% of participants was fooled by ChatGPT 3.5 in case of listening, while even 43,5% in case of reading [6][3], when trying to recognize which written composition had been produced by a human. The point, however, is not so much *if* to give confidence, but rather *how* and *why*. It

*Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy*

† These authors contributed equally. § 2 is written by E. Grande; § 3 is written by E. F. Perri.

emanuele.perri@phd.unipi.it (E. F. Perri); elio.grande@phd.unipi.it (E. Grande)

0009-0001-3906-498X (E. F. Perri); 0009-0008-2896-5900 (E. Grande)



[2] This topic was treated, *in nuce*, in [8].

[3] The quoted text will be published in the month of May or June 2024 on the journal «Mondo Digitale». We thank the authors for their courtesy.

will not be proposed here a general design model to adequately mitigate the systemic risk produced by LLMs: a too hard task. We will rather go for hunting ghosts, attempting to get closer to the nature of deception, hoping to make a little step further towards trustworthy modes of utilization of currently available devices.

## 2. *Ars Artificialiter scribendi*[4]

In *The Gutenberg Galaxy* [14], noting with Umberto Boccioni how we were (and still are, we add here) primitives of a new culture – the organic one of the electronic age which would have dulled the human consciousness in the period of its first interiorization – Marshall McLuhan remembered that the first name of the typographic printing press was "ars artificialiter scribendi" (p. 187). Weren't it for Latin, it seems coined yesterday. A way of writing, then, an art, a practical acting in the same domain of manual writing, which nonetheless had the taste of an artifice. An art of the artificial or, better, an art of elaborating a certain kind of data – in this case, alphabetic characters – in an artificial manner.

If the printing press replaced in fact the inkpot, in the corporeal movements of the hand although not in the intentions, developing LLMs is instead an "ars artificialiter scribendi" whose products appear to take up the alphabet itself, producing dialogical writing or even paradoxically oral. It would seem to be, given that we can hardly help ascribing personality, of a fine seduction strategy. Simone Natale [15] reminds us of Eliza, the chatbot invented in the Sixties by Joseph Weizenbaum, underlining the dramaturgical design, according to some "script", in the responses of new chatbots talking about trivial deceit, because it is not perceived as such and is plunged into everyday life.

However, it is not just this. Three technological breakthroughs allowed the birth of LLMs: the representation of the meaning of words through embedding, an attention mechanism to catch connections among the words themselves, and the implementation of transformers [3]. So, either some mathematics of language does exist, such that LLMs take possession of meaning – which therefore stops being «structured by fore-having, fore-sight, and fore-conception, […] the upon which of the project in terms of which something becomes intelligible as something» [10] (p. 142) – or it would regard a correlate of language itself on a parallel platform. Nothing, however, would let us think that artificial intelligence presents the fundamental property (that was) of the soul, «a being which in conformity with its kind of being is suited to "come together with any being whatsoever» [10] (p. 12), so much as the unpredictable phenomenon of the emergent abilities of LLMs. Wei et al. [21] define the "emergence" with the Nobel Prize-winner Philip Anderson as qualitative mutations in a system arising from quantitative mutations [2]. Usually, they write, scaling laws allow to foresee scale effects on systems' performance. However, at least with respect to some downstream tasks, putting the LLMs' scale on the x-axis (measured by compute, but also quantity of parameters and dataset dimension are useful indexes thereof) and performances on the y-axis, the curve does not grow up gradually but undergoes sudden variations once a certain threshold has been passed. «Note» – key point – «that the scale at which an ability is first observed to emerge depends on a number of factors and is not an immutable property of the ability». Under the category of few-shot prompting – that is, tasks apparently learned after a very small number of input instructions in the guise of teachings – comes for example the ability to reply in a truthful way or to map conceptual domains. Some performance measures, according to more than one metric, are reported by Wei et al. with respect to various typologies of LLMs (LaMDA, GPT-3, Gopher etc.), and the phenomenon of emergence appears multiple times, *but not always*, with a threshold comprised between 10^22 and 10^25 FLOPs. They are certainly tasks akin to human intellectual capabilities. However, the missing steadiness and univocity, with respect to different architectures, of the threshold to cross for an ability to emerge, lets us suspect that the emergence of new *qualities* in the behavior of such models be, yes, *correlated* with quantitative increments of compute, parameters etc., but not by them strictly *caused*. There is a semantic threshold beyond which the parts of a collection (the ancient Greek would have used here the term *pân*) are subsumed, harmonizing, in a whole (in Greek: *olòn*) where every branch, every connection finds a proper meaning. A qualitative, or at least not quantitative threshold, as it was in the sorites paradox by Eubulides of Miletus: a gap between different dimensions. It might be perhaps useful to reflect, so as to make the point on this logical mechanism, on a passage from *The Science of Logic* by G. W. F. Hegel: «Whenever all the conditions of a fact are completely

[4] Thanks to our friend Simone Farinella, PhD in history of philosophy, for the precious advice about the choice of the passage from the Hegelian work reported in this paragraph.

present, the fact is actually there; the completeness of the conditions is the totality as in the content […]. In the sphere of the conditioned ground, the conditions have the form (that is, the ground or the reflection that stands on its own) outside them, and it is this form that makes them moments of the fact and elicits concrete existence in them» [9] (p. 483). His aim was to rationalize the accidentality (nowadays we could talk about data to correlate) within unique schemes, the "things", make "real" some things which are just possible. A dimensional gap, indeed, born by the crossing of a quantitative threshold – the completeness of the conditions, which by themselves remain accidental. The problem of the representativity of data lies behind the corner.

Can an extended net of sequences, like for example the hypertext (obviously, simplifying) called "the web", overcome that critical mass and reflect, adequate itself to a systematic whole, a semantic *olòn*, a complex of signifiers? We would be tempted to reply positively: the web is our *Zeitgeist*. It contains analogies, additions in column, sentiments, errors: the patterns recognized by the emergent abilities of LLMs. Supposing to train a model – like a transformer endowed with 175 billion parameters – on such a net of sequences as dataset, won't such patterns or sub-patterns emerge? Without, among other things, real learning: the model runs in inference mode.

However, it was said that conditions – translated: correlations among data – have their ground outside themselves. The model just computes. It has only a surrogate intelligence and even a large number of parameters can't produce such improvement in quality. *But might it be good enough to mirror the improvement in quality originally lying in data semantics?* If so, we could perhaps explain why, to whom reads on the screen, a string will seem a reply, two a discourse, and a thousand a writer, although the LLM actually speaks alone, according to a hierarchy of the most probable terms.

## 3. "Somatization" of LLMs: rethinking ethics of generative AI from a phenomenological perspective

Continuing the use of the ethical-philosophical lens to study the implications of irresponsible use of LLMs (such as GPT-x, LaMDA, LLaMA, Gemini, etc.), it seems interesting and above all useful to fetch Andy Clark and David Chalmers' brilliant phenomenological formulations of the concept of extended mind [4] and Kim Sterelny's concept of scaffolded mind [20]. Thinking of *responsible LLMs* according to the standard framework (transparency, fairness, privacy, etc.), it is appropriate to ask whether a stable social trust in such technologies is not promptly impeded due to a misconception of generative artificial intelligence itself. Clark and Chalmers, in their well-known work *The Extended Mind*, bring up the example of "Otto's notebook": Otto is a patient with Alzheimer's disease who, to cope with daily mnemonic challenges, relies on a bloc-notes on which he's used to jot down and retrieve information that he is no longer aware of, due to his disease. The "analog" relationship between Otto and his notebook pours into dependence—a blind reliance; Otto's life memories are scattered around in the pages of his notebook, which is the only acceptable resource for reporting on a past and being aware of the present. The phenomenology of the notebook lies in its being much more than an external resource while retaining its original ontological status: the notebook is a *cognitive extension*, a ramification of Otto's mind and, even, a supplement to his memory. Kim Sterelny picks up on Clark and Chalmers by introducing what is a full-fledged fair corrective: the notebook, being physically *outside* the body, cannot extend cognitive capacities while also guaranteeing the same degree of reliability as the resource it replaces (that is, memory) and, therefore, its function is somewhat to support it—to *scaffold* it [20]. In other words: external (informational, datal, executive, …) resources should not be considered reliable to the same extent as internal resources since, even though external ones collaborate in dense mental associations, they are disembodied and indirectly managed. Certainly, due to mental plasticity, there are several pros of incorporating external adjuvant resources within the cognitive system—the notebook supplants memory, the cane mitigates claudication, the lens enhances vision, etc.—, but the cons, on a risk-benefit scale, are significant: (1) reliance on the external resource is inherently fallacious, since the same degree of integrity as the internal resource cannot be guaranteed; (2) exposure to the risk of sabotage of the external resource is substantial, both in the sense of environmental conditioning and in the (rarer, but not negligible) sense of targeted attacks; (3) in cases of substitution of the internal resource with an external one, an acceleration of the depletion of the already damaged internal system can be expected, causing its ultimate downfall. In this frame the relationship between internal and external environment and the environmental niche is designed—under the same risky conditions under which sentient beings gain a being-in-the-world [10]. The reflections advanced thus far soon make sense if we reimagine the

(progressively obsolescent) concept of *human-machine interaction* (HMI) from a phenomenological perspective: an environmental niche hinged on the relationship between digital system (a computer, a model, etc.) and organic system. LLMs, according to this interpretation, are the external resource—so appealing, so addictive, so affordable—with which we compensate major "*humanliest* flaws"—executory promptness, memory capacity, mundane transiency—at the risk of self-causing depletion.

Very related to this point is the risk of an only apparently reliable AI: the cognitive depletion triggered by a gradual (and not totally voluntary) renunciation of creative and cognitive capacities, which today goes hand in hand with the so-called *deskilling*; we fall into what Manfred Spitzer [18] calls *digital dementia*: an over-reliance on technology that shows potential to replace human capacities can induce a decrease in cognitive capacities for information processing and creative production (think imagination), implying symptoms close to those of dementia and that regress very slowly by suspending the use of that given technology. Spitzer writes in *Information technology in education: risks and side effects* [19] about neuroplasticity and the use of technology in learning:

«Given what we know about neuroplasticity, i.e., learning and the brain, it is hard to believe that some education practitioners and policy makers still believe that reducing cognitive load is beneficial for the learner. Quite the opposite is the case: The more effort you have to take, the better the learning outcome» (p. 84).

What Spitzer remarks is the value of direct experience, of concrete and hard doing, for a stable imprint of the information; the full experience, moreover, means taking the needed time—a permission that our postmodern society "of impatience" often does not grant. In short: doing, taking the necessary time, on the one hand; outsource for all at once, on the other. The difference between the two approaches is quali-quantitative and lies in the permanence of the result, as well as in the result itself. A similar warning comes from Stefano Cabitza who writes about epistemic sclerosis [7]:

«[...] machines AI, initially conceived to enhance peculiar capacities of men "for the benefit of men" [...], [have ended up] paradoxically to produce an opposite effect [...] of disempowerment, according to a dynamic already known to popular wisdom when it is said that "the muscle that is not used, atrophies." [...] we have called this danger "epistemic sclerosis," meaning [...] the risk of losing the habit of exploring the unknown and managing, also understood in terms of awareness, tolerance and even appreciation, the uncertainty that affects all our evaluations, estimates, predictions»[5] (pp. 80, 85).

Cabitza's is not an apologia for slow-working, nor is ours meant to be an oracle-like dystopian invective against GAI: it is, rather, about recognizing the implications of LLMs on the future of creativity, information, cultural production, and learning. Cognitive depletion [17] arises not from balanced coexistence with technology, but from *replacement by* technology, as Adriano Fabris points out at UCSI, on the topic of journalism and AI:

«[...] at best, a deskilling [...], and at worst, prospectively, a replacement of what these can do by what the AI program can do faster and more fully»[6] (§2) [5].

Just as the notebook, referred to by Clark and Chalmers, throws Otto into a relationship of absolute dependence and, virtually, worsens his memory (sparing him the stresses of exertion), LLMs, with their features simulating *Gestaltic* qualities, drag users into a relationship of dependence that affects not only the most time-consuming mechanical activities, but also the most human and light ones (drafting an e-mail, replying to a message, ...); what are the long-term effects of such a dependence of this extent? At the beginning of the paragraph we made a reference to the fundamental unacceptability of the external resource when it has function of cognitive extension, given three key cores; those same three cores can be repurposed to contribute to a new framework for responsible and reliable GAI; in the present case, for example, considering a multimodal transformer as an external resource (with a function of cognitive extension that is, extended mind), it will, if heavily used, necessarily have to produce adjuvant effects—it will be notebook, will be cane, will be lens, …—and other "castrating" ones: (a) in being an external resource, it will not guarantee continuous accessibility, (b) it will be subject to environmental conditioning or manipulation—especially since datasets are generally neither personal nor personally inspectable/customizable (except for sparse

[5] English translation provided by the authors.

[6] English translation provided by the authors.

instances of *RLHF* like temporary slight changes in model behavior based on user-expressed preferences via *A/B testing*) —, (c) it will worsen cognitive capabilities, which are already compromised [13] and there will be instances of outright dependency. It is evident, as the last decades of pocket electronics, phenomenology and philosophy of mind teach (also showing us several cases of so-called *adaptive phenotypic plasticity*), that whatever technology shows the prerequisites for cognitive extension is in the long run pejorative of cognitive abilities and, by extension, of the-being-in-the-world respecting the physiological sharing/reserve alternation. In order to build lasting social trust and ensure a healthy coexistence with generative AI and whatever other technology will result—this is also the EU's approach[7] [1]—it's crucial to talk about ethics: while it is necessary to ensure an ethics *in* AI, it seems more important to work on an ethics *of* AI: introducing the teaching of ethics (in general) and AI ethics as early as K-12 [12] is the only way to lay the groundwork for a truly accountable and reliable GAI. Admittedly, the utterly interdisciplinary nature of such an endeavor is well-known by this time; it remains, however, that ethics and law are the only two cartridges to foster the desired healthy coexistence. Given the "position paper" nature of this contribution, it is worth repeating that the writers' intent is to emphasize the importance of introducing ethics from the earliest years of schooling: at stake is the replacement of human creativity with generative sterility resulting from the statistical prediction of language—*P=(W|h)*, if we talk about word-embedding in NLP—, to disrupt not only the field of culture, but also the very criteria of aesthetic-artistic evaluation of written opuses. A separate parenthesis is to be opened in regard of biases management in generative AI—a hot topic in the area of *responsible AI* practices. "GAI bias" means the systematic trend of a generative model to return outputs biased toward certain responses; the reasons why this happens can be attributed to the dataset used for training, to implicit assumptions during the training itself, or even to biases inherent in our society and thus reflected in the "answers" given by the system.

That of bias in transformers is often considered a problem that we still need to solve interdisciplinarily, a problem that undermines the path to "responsible and reliable" GAI. The feeling is that we cannot see the wood for the trees: the problem lies elsewhere, outside the development and usage patterns of AI systems; the biases are in the training data since they mirror what our society has produced to date. To put it another way: writing a prompt to a chatbot asking for the writing of a text à la D.A.F. de Sade and ending up complaining about a bias for the degrading representation of women versus that of a violently dominant man is laughable. It would seem right, somewhat, to accept the biases for what they are: reflections of what we have been; then, a GAI is all the more reliably "responsible and trustworthy" when it transparently represents a state of affairs, not when it works of embellishment. The new front in the struggle for transparent AI is demystifying the fight against bias; it has to do with the exercise of moral posture, with confrontation (even unpleasant, so be it), with history and *characterial ideal types* [11]—in the Weberian sense of simplified idealization. While the difference between character ideal type, persona (as a unique combination of attributes defining a certain individual), figural restitution and bias is *sub sole*, it is not as clear (to many AI ethicists, but not only) that the goals of transparency and trustworthiness are not pursuable by purging bias: only a generalized sensitivity to the use and consequences of generative systems will be able to avert the big issues on the horizon.

## 4. Conclusion

This paper has sought to explore the social side of reliability and accountability with respect to the use of large language models, providing a qualitative and semantic reading of the origin of the so-called "emergent abilities" of such generative models. The analysis was supported by parallels between extended mind and AI-based transformers, winking at a more phenomenological approach to the problem of GenAI misuses. Even if for a few lines only, we went "ghost hunting" motivated to investigate in the nature of these systems neither more nor less than what they are.

## Acknowledgement

"FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the

[7] «L'approccio etico dell'Unione europea alla intelligenza artificiale è volto a sollecitare una riflessione etico-umanistica sul progresso tecnologico mondiale». (Alpini, 2019, 6); Transl. by the authors: «The European Union's ethical approach to artificial intelligence is intended to prompt ethical-humanistic reflection on global technological progress».

European Commission under the NextGeneration EU programme, PNRR.

## References


[1] A. Alpini, Sull'approccio umano-centrico all'intelligenza artificiale. Riflessioni a margine del "Progetto europeo di orientamenti etici per una IA affidabile", «Comparazione e diritto Civile», 2019, 2, 1-9.

[2] P. W. Anderson, More is different: Broken symmetry and the nature of the hierarchical structure of science, «Science», 1972, 177(4047):393–396. http://www.lanais.famaf.unc.edu.ar/cursos/em/Anderson-MoreDifferent-1972.pdf (accessed 21/04/2024).

[3] G. Attardi, Il Bello, il Brutto e il Cattivo dei LLM, «Mondo Digitale», 2023, June, 1-16.

[4] A. Clark, D. Chalmers, The extended mind, «Analysis», 1998, 58(1), 7-19.

[5] A. Fabris, Giornalismo e intelligenza artificiale: la questione etica di cui parla Adriano Fabris. Unione Cattolica della Stampa Italiana, 10/02/2024. URL: https://www.ucsi.it/news/opinioni/14595-giornalismo-e-intelligenza-artificiale-la-questione-etica-di-cui-parla-adriano-fabris.html (accessed 21/04/2024).

[6] A. Fabris, P. Ferragina, I. Horvat, D. Morelli, G. Prencipe, Filosofia interroga Arte, Drammaturgia sfida IA. Due testi, due podcast, per rispondere alla domanda: scrittura umana o artificiale?, «Mondo Digitale», 2024 [forthcoming].

[7] L. Floridi, F. Cabitza, *Intelligenza artificiale: L'uso delle nuove machine*, Bompiani, Milano 2021.

[8] E. Grande, LLMs: il surrogato dello Spirito del mondo, «Fondazione Leonardo – Civiltà delle Macchine», 18/01/2024, https://www.civiltadellemacchine.it/it/news-and-stories-detail/-/detail/llms-surrogato-spirito (accessed 19/01/2024).

[9] G. W. F. Hegel, The Science of Logic, transl. G. Di Giovanni, Cambridge University Press, 2010.

[10] M. Heidegger, Being and Time. A translation of Sein und Zeit, transl. J. Stambaugh, State University of New York Press, 1996.

[11] B. Hibou, M. Tozy, Ragionare per idealtipi. Comprendere con Weber lo Stato contemporaneo in Marocco… e altrove, «Cambio. Rivista sulle Trasformazioni Sociali», 2021, 10(20), 65-83.

[12] I. Lee, S. Ali, H. Zhang, D. DiPaola, C. Breazeal, Developing middle school students' AI literacy, in Proceedings of the 52nd ACM technical symposium on computer science education, 2021. pp. 191-197.

[13] L. A. Manwell, M. Tadros, T. M. Ciccarelli, R. Eikelboom, Digital dementia in the internet generation: excessive screen time during brain development will increase the risk of Alzheimer's disease and related dementias in adulthood, «Journal of Integrative Neuroscience», 2022, 21(1), 028.

[14] M. McLuhan, La Galassia Gutenberg. Nascita dell'uomo tipografico, transl. S. Rizzo, Armando Editore, Roma 1976.

[15] S. Natale, Macchine ingannevoli. Comunicazione, tecnologia, intelligenza artificiale, transl. D. A. Gewurz, Giulio Einaudi Editore, 2022.

[16] Parlamento Europeo (2024). Emendamenti del Parlamento Europeo alla proposta della Commissione. Regolamento (UE) 2024/… del Parlamento Europeo e del Consiglio del ... che stabilisce regole armonizzate sull'intelligenza artificiale… (legge sull'intelligenza artificiale), (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), 06/03/2024, https://www.europarl.europa.eu/doceo/document/A-9-2023-0188-AM-808-808_IT.pdf (accessed 21/04/2024).

[17] E. F. Perri, Generative artificial intelligence and creative-cognitive depletion: an ethical issue. Use and abuse of GAIs and GPTs in the field of culture and education. IA, educación y medios de comunicación: modelo TRIC, Dykinson S.L., Madrid 2024, (preprint).

[18] M. Spitzer, Demenza digitale. Come la nuova tecnologia ci rende stupidi, Corbaccio, 2013.

[19] M. Spitzer, Information technology in education: Risks and side effects, «Trends in Neuroscience and Education», 3(3-4), 2014, 81-85.

[20] K. Sterelny, Minds: extended or scaffolded?. «Phenomenology and the Cognitive Sciences», 9(4), 2010, 465-481.

[21] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D: Metzler, E.H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean., W. Fedus, Emergent Abilities of Large Language Models, in Transactions on Machine Learning Research, August 2022, arXiv:2206.07682v2 [cs.CL].