AI beyond ChatGPT

Artur d'Avila Garcez

May 20, 2025

Abstract

Two years on from the release of GPT4, the fierce debate that ensued over the risks of AI has cooled off, Big Tech companies released various proprietary and open-source competitors to ChatGPT, and the European Union passed the regulatory AI Act in record time. Leading figures disagreed on what needed to be done: some claimed that Big Tech was best placed to take care of AI safety, others argued in favor of open source, and others still called for immediate worldwide regulation both of AI and social media. As society contemplated AI's impact on everyday life - with hundreds of millions of users of large language models taking part in what feels like the world's largest experiment - the technological innovations that led to GPT4 and its successors started to receive less and less attention. But the technology is central to the study of the risks and opportunities of AI. The opacity surrounding AI since the release of ChatGPT contributes to fears of existential risk and fuels the claims of an upcoming AI bubble burst. Without a clear understanding of the technology and Big Tech's use of data, regulatory efforts are in the dark. In this opinion article, I seek to refocus attention on the AI technology's achievements and limitations. I argue that the emerging field of neurosymbolic AI may address the problems of current AI: lack of fairness, reliability, safety and energy efficiency. I address the issue of accountability in AI as a topic that is broader than regulation, and point out that so-called frontier AI is high-risk, not because AI may take over humanity, but because of a lack of accountability in the face of large-scale spreading of misinformation.
Keywords: Generative AI, Accountability, Neurosymbolic AI.

1 Five days of chaos

The New York Times, which is now suing GPT4's owner OpenAI for copyright violation, best summarized the drama that unfolded at OpenAI after the release of GPT4 as *five days of chaos*[1]. An OpenAI board member had written an article stating "OpenAI has also drawn criticism for many other safety and ethics issues related to the launches of ChatGPT and GPT4, including regarding copyright issues, labor conditions for data annotators, and the susceptibility

[1] `https://www.nytimes.com/2023/11/22/technology/how-sam-altman-returned-openai.html`

of their products to jailbreaks that allow users to bypass safety controls" [3]. Following the departure of key technical talent, some of whom went on to create an AI safety company, little seems to have changed in OpenAI's governance or accountability approach. ChatGPT's lack of reliability, fairness and data efficiency has been noted many times since, e.g. [12], with perhaps the most prominent example being the case of a law professor falsely accused of sexual harassment. I will argue that this kind of problem cannot possibly be solved after the event and case-by-case, although after seeing so many such cases we all now seem to have become desensitized to the problem. There must be a better way of achieving AI that can offer certain guarantees to model alignment with a lower financial and human cost. Next, I will emphasize the need for an accountability ecosystem as proposed in [10]. I will focus on how technology can be leveraged to promote accountability in AI. A lot of the technological claims from two years ago hinged on RLHF (Reinforcement Learning with Human Feedback). We now know that RLHF is both unethical, exposing data labelers to the worst of the internet, and too costly. Some of the competition, DeepSeek, adopted a different approach based on *distillation*, something that I will discuss later and that is closer to the neurosymbolic approach.

2 Risk and Accountability in AI

Large Language Models (LLMs), including OpenAI's ChatGPT, Meta's Llama and Google's Gemini, can all be viewed as very large computer programs. The programs are learned from examples (inputs and their desired outputs mapped to vector representations given very large amounts of text data scraped from the internet). The learned program function f is capable of producing coherent sentences in response to human interaction (prompts). LLMs produce answers - right or wrong - to any given question. They show reasoning capabilities and can offer explanations to the answers and, most impressive of all, can produce code in a programming language given prompts in natural language. LLMs are a great engineering achievement, are excellent at text summarization and language translation. They may help improve productivity of anyone who is diligent and sufficiently knowledgeable to check when the LLM might have made a mistake, and yet they have great potential to deceive all those who are not diligent or sufficiently knowledgeable. As an auto-regressive model, LLMs do everything they do by doing only one thing: predicting the probability of the next word token x_{t+1} in a sequence of words, given the current tokens x_t at time t, where $x_{t+1} = f(x_t)$. Having made a choice of the next token, LLMs will apply the same program function f again, recursively, to build a sentence with many more word tokens.

Should tech companies have been allowed to release LLMs worldwide to hundreds of millions of users and collect their data without any external scrutiny? There are various technical and non-technical reasons why current AI, based on large-scale transformer neural networks, may not be deployed in this way: lack of trust or fairness, reliability issues and public safety (as in the case e.g. of

self-driving cars using the same large scale neural network models). Fixing reliability issues case-by-case with RLHF has proven to be too costly, financially and in human terms, as already pointed out. A commonly-adopted risk mitigation strategy became known as the *human-in-the-loop* approach: making sure that a human is ultimately responsible for any decision making. In such cases, the AI system is seen only as an assistant to the humans who are supposed to be in control. However, the situation is more nuanced. Simply apportioning blame or liability to humans does not solve the problem. It is necessary to empower the user of AI, the data scientist and the domain expert, to interpret, ask *what-if* questions, and if necessary intervene in the system. Here, the technology of explainable AI becomes key. Having very large and opaque LLMs doesn't empower the user. Consider LLMs' ability to produce code. If ChatGPT was allowed to work, not as a stand-alone computer program function f, but in a loop whereby the code that is generated can be executed automatically, data collected from the execution, and function f changed to seek to improve the code, one can see straight away how such self-improving LLM with decision making autonomy - termed agentic AI - may pose a serious risk to current computer systems. Recent experiments indicated that at present this loop of automatic synthetic data generation can create the opposite effect, self-impairing, producing a degradation in performance [11]. Agentic AI, therefore, will require guardrails once it is allowed to take action on one's behalf, even if that action is something as simple as organizing one's holiday, paying for the air tickets and deciding on the sightseeing tour schedule. The idea of a self-improving loop blurs the distinction that exists currently between model training and test-time compute. Up until now, the separation between training and run-time has been a very useful tool in the engineering of large AI models.

3 AI challenges: reasoning, data efficiency, fairness and safety

When LLMs make stuff up such as referring to non-existing articles in the above-mentioned case of the law professor, they are said to hallucinate. AI will require systems that never hallucinate, that reason reliably, that can handle novelty and treat exceptions requiring fewer data labeling. This is very different from the current *scale-is-all-you-need* LLM approach. Two years on, LLMs continue to hallucinate even with the costs of post-hoc model alignment with RLHF skyrocketing as performance seems to plateau on benchmarks. Take for example OpenAI's o3 LLM system. It was claimed to "think before it answers" and to be capable of "truly general reasoning". Little is known about the inner workings of o3, but it is reasonable to assume that it works as a kind of "GPT-Go": a pre-trained transformer to which test-time compute is incorporated in the form of a search process in the style of Google DeepMind's earlier Alpha-Go system. The search uses "Chain of Thought" (CoT) prompting: generation of synthetic data using the transformer itself in a chain that breaks down a prompt into sub-

prompts. It is reasonable to expect an increase in test-time compute to improve reasoning performance because reasoning tasks are typically solved by breaking down a problem into sub-problems. In fact, neurosymbolic approaches have shown improvement in reasoning performance when learning is combined with search-based reasoning. In [7], knowledge distillation is used to build a search tree from a trained neural network, with the search tree used to carry out formal reasoning at test time. The problem with the "GPT-Go" approach is the lack of reliability of the synthetic data generation, known as the curse of recursion [11], and the combinatorial nature of the CoT input, best described as *infinite use of finite means* (a finite dictionary giving rise to infinite possible texts), that is, the well-known problem that small changes in input may produce diverging outputs due to the inevitable accumulation of errors in the calculations of a neural network. As a result, CoT may solve one reasoning task today, only to fail at an analogous reasoning task tomorrow. This is best illustrated by the examples provided in [6], showing that a mere change in naming convention can affect reasoning performance dramatically. What we see in practice is that eliminating so-called hallucinations is very hard, if not impossible.

In neurosymbolic AI, based on a long tradition of combining neural networks with symbolic knowledge, the goal is to ascribe meaning to neural computation by offering a formal semantics to neural networks. Instead of merely adjusting the inputs of the network (the CoT approach), the neurosymbolic approach designs the network architecture or the training loss function based on symbolic descriptions that are either learned or already known. The intended results are more confidence in the outputs of the network and a more parsimonious learning as a result of improved modularity.

Neurosymbolic AI integrates learning and reasoning as part of model development by following a development cycle known as the neurosymbolic cycle: (i) extract symbolic knowledge descriptions from partially trained networks, (ii) reason formally about what has been learned, (iii) compress the network by instilling consolidated knowledge back into the network before further training with data. Reasoning in neurosymbolic AI follows the tradition of knowledge representation in AI, founded on logic and formal definitions of a semantics for deep learning [9], rather than based on informal evaluations of reasoning capabilities using benchmark data. Evaluating neural networks with respect to formally-defined, sound or approximate reasoning allows for the much needed controlling of the accumulation of errors. The use of distillation is a step in the direction of neurosymbolic AI in that distillation is a form of knowledge extraction to obtain network compression (steps (i) and (iii) of the neurosymbolic cycle). The use of test-time compute is also a step in the direction of neurosymbolic AI because reasoning is, in essence, implemented in a computer by means of a search process (step (ii) of the neurosymbolic cycle). However, there are many forms of knowledge representation and reasoning to be mapped onto neural network models (analogy, modal, epistemic and higher-order logic, temporal, normative, abductive and abstract reasoning, etc.) and distillation without explicit knowledge, that is, description and semantics, is limited in what it can offer to reliability, explanation, knowledge reuse and transfer learn-

ing. Broadly speaking, neurosymbolic AI is tasked with the theory, algorithms and tools capable of establishing a principled understanding of the correspondences that exist between neural and symbolic representations. Ultimately, the intended outcome of neurosymbolic AI is to achieve safety via verifiable descriptions of network modules, energy efficiency via parsimonious learning from data and knowledge, fairness by imposing requirements specified in logic, and trust by empowering the users of AI with the help of explainability.

4 Desiderata for neurosymbolic AI

Addressing the challenges of data efficiency, safety, fairness and trust in AI will require breakthroughs in neurosymbolic AI. Neural computation has shown with deep learning that it must be the substrate of AI, the foundation layer upon which AI is implemented, but each of the above problems with deep learning have been stubbornly difficult to fix. Already in the late 1990's and early 2000's, the importance of neural computation as the substrate of AI was obvious to a small group of researchers advocating neurosymbolic AI. The value of symbol manipulation and abstract reasoning offered by symbolic logic was also obvious to that group. It could be argued, however, that neurosymbolic AI starts together with connectionism itself, with McCulloch and Pitts's 1943 paper *A Logical Calculus of the Ideas Immanent in Nervous Activity*, which influenced John Von Neumann's 1952 *Lectures on Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components*, both indicating that the separation between distributed vector representations (network embeddings) and localist symbolic representations (logic) did not exist. Even Alan Turing's 1948 *Intelligent Machinery* introduced a type of neural network called a B-type machine. It was not until after the term *Artificial Intelligence* was coined by John McCarthy, admittedly for the sake of securing funding ahead of the now famous Dartmouth Workshop in 1956, that the field separated into two: symbolic AI and connectionism (or neural networks). This has slowed down progress in AI as the two research communities went separate ways with their own conferences, journals and associations. Following the success and downfall of symbolic AI in the 1980's (the first wave of AI) and the success since 2015 of deep learning with its now obvious limitations around fairness, safety, energy efficiency and a lack of explainability (the second wave of AI), neurosymbolic AI is regarded by many as *the third wave of AI* [1].

Neurosymbolic AI uses knowledge extraction to explain and, if needed, intervene in the AI system. The goal is to control the learning process in ways that can offer correctness or fairness guarantees to the neural network, producing a more compact and hopefully more efficient network in the process. The 2024 Chemistry Nobel prize-winning AI system *AlphaFold* from Google DeepMind is arguably the greatest achievement of AI to date. AI systems of this kind (in the case of *AlphaFold*, a system for protein structure prediction) hold the promise to cure many diseases. They are application-specific systems also known as *narrow AI* compared to LLMs. From particle physics to drug synthesis, energy

efficiency and novel materials, AI is being adopted as the new language of scientific discovery. However, failure to provide a description capable of conveying a deeper sense of understanding of the solutions being offered by AI can be very unsatisfactory. In a closed environment, such as a board game, large-scale simulation may be sufficient to get the network to learn to reason. In open-ended situations, the reasoning task is much harder than reasoning by similarity, requiring the use of an explicit description. An explicit description is one that can be manipulated by asking the question *what might happen if I were to make this change?*, without making the change. Hence, the description is required to be amenable to symbolic manipulation. AI will soon require systems that adapt to novelty from only a few examples, that check their understanding, that can multi-task and reuse knowledge to improve data efficiency and that can reason in sophisticated ways using first-order and higher-order logic. Adapting to novelty requires an ability to create abstract, simple representations (whether in the brain or the mind) but also to change representations from time to time [4]. Change of representation allows looking at a problem from a different angle to obtain new insight given an analogous situation. This forms the core challenge of the latest research in neurosymbolic AI: (i) extraction of relevant descriptions from complex, very large networks at an adequate level of abstraction for the application at hand, (ii) sound reasoning and learning with various representations: spatial, temporal, abstract, relational and multimodal [2], (iii) data and knowledge reuse with modularity to extrapolate efficiently to multiple tasks in different domains of application. The next decade of research in neurosymbolic AI should make key strides towards addressing this challenge.

5 Avoiding a race to the bottom

Now that the AI race is on, influential leaders have been arguing for more investment in safety research or industry regulation. It should be obvious that worldwide regulation is not achievable in the current geopolitical situation [5]. An alternative to regulation has been put forward in [10]: digital technology itself, as part of an adequate AI accountability ecosystem, can offer a new path to a more parsimonious AI where neural models are more modular, trained with fewer data and validated symbolically during multiple stages of model development. This is quite different from what the EU AI Act has achieved. Regulation without accountability increases risks by encouraging weak competitiveness.

A proper accountability ecosystem can avoid a race to the bottom by mapping out general principles into implementation of industry processes using mechanisms such as internal auditing, external accreditation, investigative journalism, a risk-based regulation approach and market shaping. In [10], already in 2021, it was stated: "at present the ecosystem is unbalanced, which can be seen in the failures of certain mechanisms that have been attempted by leading technology companies. By taking an ecosystem perspective, we can identify certain elements that need developing for the system as a whole to function effectively. Corporate governance mechanisms such as standardized processes

and internal audit frameworks, leading to potential external accreditation, need to be made to work together in ways that go beyond regulatory requirements, especially in technologies' early period of evolution and deployment when regulation lags practice." Key stakeholders in this process include corporate actors, market counterparts, academia, civil society and government.

An AI system to predict harm from online gambling was used as a case study in [10] because of the high regulatory focus, divergent regulatory perspectives worldwide and longstanding debates over ethical dilemmas in gambling. Two key elements of the accountability ecosystem were discussed: (i) interventions to reduce bias and (ii) increased transparency via model explainability. The benefits of having an industry-specific accountability process were illustrated to the extent that it can be documented, reviewed, benchmarked, challenged and improved upon, both to build trust that the underlying ethical principle is being taken seriously and to identify specific areas to do more. Results were drawn from the risk profiling of gambling behavior: symbolic knowledge was extracted from a neural network predicting problem gambling and the explainable AI (XAI) technology was evaluated on indirect gender bias and algorithmic fairness metrics. It was argued that effective regulation requires accountability, the adoption of a risk-based approach, and the definition of a risk-mitigation strategy informed by objective metrics, such as fidelity of knowledge extraction when mapping neural and symbolic processes with the use of XAI [13].

6 Conclusion

I argued that achieving accountability in AI with data efficiency, fairness and safety will require a new approach based on neurosymbolic AI. Although I am convinced that neurosymbolic AI can make AI fairer, safer and more energy efficient, it is difficult to see how the use of technology alone could solve the problem of misinformation at scale polluting the internet with unreliable AI-generated data. Although the widespread adoption of AI technology should be celebrated in its potential to increase productivity, current GPT-style AI will also magnify errors as users become complacent. Malicious users have been spreading misinformation at a fast pace while rushed regulation without accountability has failed to protect the public. In particular, lessons were not learned from the failures to protect children from harm in the context of social media.[2] As humanity has to adapt on-the-fly to *the big LLM experiment*, AI-based errors and misinformation seem to have to be accepted as a "fact of life". Controlling the large-scale spreading of misinformation is no longer in the command of any social media platform and the cost of checking for misinformation is increasing considerably. At the heart of the problem is a business model that has caused turmoil in corporate media, where software is perceived to be free, when in fact it is paid for with data. Alternatives are needed to this business model while concerns around freedom of speech, the inadequacy of current incentives

[2] `https://www.commerce.senate.gov/2021/10/protectingkidsonline:testimonyfromafacebookwhistleblower`

and other moral dilemmas prevent companies and governments from making changes.

One idea is to treat data as fiduciary money, the idea of *data banks*, allowing users to pay for bits consumed, in exchange for privacy, while paying users for bits produced. For this to work, payment systems would have to scale-up to allow for micro-payments to take place at close-to-zero transaction cost, while addressing the problem of unique digital identity. Given a choice, people may prefer, differently from the subscription model, to *pay-as-you-go* for bits of (add-free) information. As AI moves from the academic labs into everyday life, new ways of doing the things that we take for granted will need to be decided upon and implemented quickly. Mechanisms of technology-enabled local direct democracy may help communities manage this transformation [8].

AI is not only changing the world of employment, but also education. Learning at schools and universities will need to change to instill a culture of critical and creative thinking, of learning from the history of science, and to cultivate the values of scientific inquiry, promoting skeptical interrogation based on sound principles of uncertainty quantification. AI will need to be taught at schools with the goal of creating a more discerning and informed society. Leaders, decision makers and domain experts should probably also learn the basics of AI. It has taken society many years to learn to distinguish genuine from malicious websites. Learning whether or not to trust the output of LLMs is much harder and will require the help of technological advancements such as explainable and neurosymbolic AI, but also a new economic and social contract, empowering local democracy, requiring fast policy decisions in a very fast-changing world.

Acknowledgments: I thank Chris Percy, Simo Dragicevic, Luis Lamb and Moshe Vardi for valuable discussions on accountability in AI and for their comments on earlier versions of this article.

References

[1] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *Artif. Intell. Rev.*, 56(11):12387–12406, 2023.

[2] Artur d'Avila Garcez, Luis C. Lamb, and Dov M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Springer, 2008.

[3] Andrew Imbrie, Owen Daniels, and Helen Toner. Decoding intentions. `https://cset.georgetown.edu/publication/decoding-intentions/`, October 2023. Center for Security and Emerging Technology [Online; accessed 20-Jan-2025].

[4] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.

[5] Chris Miller. *Chip War: The Fight for the World's Most Critical Technology*. New York, Scribner, 2022.

[6] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024.

[7] Kwun Ho Ngan, James Phelan, Esma Mansouri-Benssassi, Joe Townsend, and Artur S. d'Avila Garcez. Closing the neural-symbolic cycle: Knowledge extraction, user intervention and distillation from convolutional neural networks. In *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, Siena, Italy, July 3-5, 2023*, volume 3432 of *CEUR Workshop Proceedings*, pages 19–43. CEUR-WS.org, 2023.

[8] Ritesh Noothigattu, Neil Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proc. AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1587–1594. AAAI Press, 2018.

[9] Simon Odense and Artur d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artif. Intell.*, 340:104273, 2025.

[10] Chris Percy, Simo Dragicevic, Sanjoy Sarkar, and Artur d'Avila Garcez. Accountability in AI: From principles to industry-specific accreditation. *AI Commun.*, 34(3):181–196, January 2021.

[11] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross J. Anderson. The curse of recursion: Training on generated data makes models forget. *CoRR*, abs/2305.17493, 2023.

[12] Benedikt J. Wagner and Artur d'Avlia Garcez. A neurosymbolic approach to AI alignment. *Neurosymbolic AI journal*, 1:1–12, August 2024. IOS Press.

[13] Adam White and Artur S. d'Avila Garcez. Measurable counterfactual local explanations for any classifier. In Giuseppe De Giacomo et al., editor, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 Aug - 8 Sep 2020, Santiago de Compostela, Spain*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2529–2535. IOS Press, 2020.