Helping AI grow up without pressing 'pause'

Edited by:
Adam Leon Smith FBCS - Chair of BCS, The Chartered Institute for IT Fellows Advisory Group (F-TAG) and CTO Dragonfly; and Dr Sam De Silva CITP FBCS - Chair of BCS Law Specialist Group and Partner at international law firm CMS Cameron McKenna Nabarro Olswang LLP.

Contributions by other F-TAG members and BCS Specialist Group members listed below

Summary

This paper argues that a pause in the development of 'more powerful' AI as called for in an open letter by the Future of Life Institute (29 March 2023) is not the most practical way to manage the risks of AI development.

BCS, The Chartered Institute for IT argues that a pause is likely to:

Not work in practice and be unrealistic.

Delay developments in AI research in areas that are crucial for solving society's problems in relation to issues such as climate change and diagnosing and treatment of disease.

Result in a position which is 'asymmetric', in that it is not possible to ensure all governments and organisations would comply with a request for a pause.

This could provide bad actors an advantage in developing AI for nefarious purposes.

[1] **Other BCS F-TAG members contributing:** Professor Tom Crick MBE FBCS; Professor James Davenport FBCS FIMA CITP CMath; Hema Purohit FBCS; Professor Mark Josephs FBCS CITP; Christine Ashton FBCS; Professor Margaret Ross MBE FBCS; Max Hemingway FBCS CITP. **Other BCS members contributing:** Professor Adrian Hopgood FBCS (BCS AI Specialist Group); Steve Sands FBCS (BCS Information Security Specialist Group)

We argue that instead AI can and should continue to develop, whilst risks are managed as far as practicable, which means:

- Organisations are transparent about their development and deployment of AI, comply fully with applicable laws (e.g. in relation to data protection, privacy and intellectual property) and allow independent third parties to audit their processes and systems.
- There are clear and unambiguous health warnings, labelling and opportunities for individuals to give informed consent prior to being subject to AI products and services.
- AI is developed by communities of competent, ethical, and inclusive information technology professionals, supported by professional registration.
- AI is supported by a programme of increased emphasis on computing education and adult digital skills and awareness programmes to help the general public understand and develop trust in the responsible use of AI, driven by government and industry.
- AI is tested robustly within established regulatory 'sandboxes' as proposed in the white paper to guide the use of AI in the UK published by the Department of Science, Innovation & Technology on 29 March 2023 (the 'AI White Paper').
- The use of sandboxes should be encouraged beyond a purely regulatory need - for example to test the correct skills and registration requirements for AI assurance professionals and how best to engage with civic societies and other stakeholders on the challenges and opportunities presented by AI.

Background

Generative AI, and particularly large language models (LLMs) have gained a lot of attention in recent months, triggered largely by the public release of OpenAIs ChatGPT, closely followed by Google's Bard and Microsoft's Bing.

In essence LLMs are text processing software that can generate text that appears almost identical to that which would be generated by a human. They appear to be impressive in their capability to gather data from the internet and discussions, then piece it together in response to queries. Unfortunately, there are a number of issues with LLMs. As widely reported, they are often inaccurate, in part because they prioritise generating convincing text over accurate text. They are not sentient and accordingly not aware of what the text they are generating means. In other words, the machine learning models in LLMs understand relationships between words, not relationships between the meaning of words. LLMs may also perform poorly at other tasks, such as basic maths, because they are a narrow and specific application of AI.

The application of LLMs uses a complementary form of AI known as knowledge-based or symbolic AI. These techniques have their origins in the 'expert systems' that were popular in the 1980s and 1990s. Such systems allow the human expertise that has been acquired to be applied in a range of unpredictable scenarios. One example of this would be to ensure that the LLMs output is not obviously discriminatory or breaching any laws. While of course this is not a bad thing, it masks the shortcomings of the underlying machine-learning algorithms. It creates an illusion that issues are less present than before, when in fact they are simply concealed.

[2] BCS Webinar, https://www.youtube.com/watch?v=lSwKtOZ15ew

[3] G. Marcus and E. Davis. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about: Tests show that the popular AI still has a poor grasp of reality. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion,2020.

[4] Hopgood, A.A. Intelligent Systems for Engineers and Scientists: A Practical Guide to Artificial Intelligence, 4th edition. CRC Press, (2022), 451 pages, ISBN 9781032126760.

[5] Weizenbaum, Joseph (1976). Computer Power and Human Reason: From Judgment to Calculation. New York: W. H. Freeman and Company. ISBN 0-7167-0464-1.

In a sense, we view ChatGPT (and other LLMs) as a distraction in relation to increasing the reasoning power of AI systems. While LLMs are AI systems that can generally respond to human queries well, they do not represent a significant improvement in reasoning ability. We acknowledge that LLMs do appear to be able to reason better on first glance, but that is because they have illusory anthropomorphic qualities. The tendency of humans to anthropomorphise should never be underestimated. The earliest examples of human behaviour exhibited by machines had this effect, these were not AI as we understand it today, but merely crude pattern matching.

The problems with such systems are not always evident in common use, but in one example, GPT-3 has been found to advise patients with mental health issues to die by suicide. Many other examples of unsafe responses can be found and envisaged, and it will be impossible to prevent these without significant development in the state of the art. This is even acknowledged by GPT-3 itself, which says *"... GPT-3, like other language processing models, can generate biased or offensive text if it is trained on biased data. This can be a problem if the generated text is used in sensitive applications, such as in healthcare or education, where it is important to avoid bias and promote inclusivity"*.

We are aware that some experts are recommending the UK builds its own sovereign LLM. The Alan Turing Institute is also recommending that there is some foundation model capability established in the UK. Whilst clearly the UK must continue to build its AI capability, we urge careful strategic thought about what building a sovereign LLM capability actually means.

Since launching to the public in late November 2022, ChatGPT reached 100 million active users in just two months, making it the fastest growing consumer application in history. However, the significant use by end-users has been accompanied with significant concerns from some AI experts and regulators.

For example:

- On the 29th March 2023, the Future of Life Institute issued an open letter calling for a pause in the development of 'more powerful" AI while guard rails could be put in place.
- On the 30th March 2023, the US Centre for Artificial Intelligence and Digital Policy submitted a complaint to the US Federal Trade Commission (FTC). The complaint outlines many concerns about GPT and requests an investigation, and further a halting of further commercial deployment; independent assessment of further GPT products prior to deployment; and a publicly accessible incident reporting mechanism.
- On the 31st March 2023, the Italian privacy regulator ordered ChatGPT to stop processing Italian citizen's data and issued an enforcement notice to OpenAI. The regulator said that OpenAI lacks a legal basis justifying 'the mass collection and storage of personal data ... to "train" the algorithms' of ChatGPT. OpenAI also processes data inaccurately, it added.
- On the 4th of April 2023 the Office of the Privacy Commissioner of Canada said that it was investigating ChatGPT after it received a complaint claiming the software was collecting, using, and disclosing personal information without consent.

[6] R. Daws. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-, 2020.
[7] F. Sowa. [ontolog-forum] fw: What GPT-3 "thinks" of GPT-3. E-mail to Ontology Forum 10 December 2022, 2022.
[8] https://techmonitor.ai/technology/ai-and-automation/ai-uk-large-language-model-chatgpt
[9] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01
[10] https://www.theguardian.com/technology/2023/mar/29/elon-musk-joins-call-for-pause-in-creation-of-giant-ai-digital-minds
[11] Some of the researchers cited in the letter disagree with it - https://www.reuters.com/technology/ai-experts-disown-musk-backed-campaign-citing-their-research-2023-03-31/
[12] https://www.caidp.org/cases/openai/
[13] BCS podcast - https://lnkd.in/eFKEZVQb
[14] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847
[15] https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230404/

Should we pause AI development?

While there are clearly significant issues that need to be considered in relation to LLMs, we consider that pausing all AI research and development, as suggested by the Future of Life Institute open letter, would delay innovation in areas that are far more fundamental for humanity than text generation, including climate change and hunger.

While significant advances have clearly been made in LLMs, it is not entirely clear that this rapid pace of development will continue, although we expect that LLMs will become more integrated with other tools and become increasingly multi-modal (for instance, consuming and generating images and sound in addition to text). We believe that the idea that 'too powerful' AI is going to emerge in the next six months, or even six years, is somewhat misleading. Most of the work that is having an impact is specific AI, which is narrow in its application.

The concerns from most experts are not that AI is too powerful, but that basic guard rails are not in place to ensure AI is deployed responsibly. It should be noted that the guard rails that need to be put in place are governance-based rather than being technical innovations. They could be commercial and operational in nature, for example technology providers providing more transparency about the AI systems they are providing. Also, such technology providers could also allow independent audit of their processes and systems.

AI is expected to deliver significant benefit for the UK population, it is booming in healthcare and biology, and it can help us mitigate the impact of climate change. A pause on these critical innovations could be harmful in these and other such critical areas.

Another issue with a pause is that it is likely to be asymmetric. Companies, researchers, organised crime groups and entire nations (including the UK) may decide not to pause, for reasons already mentioned, or in order to gain advantage. Accordingly, we think a pause is unlikely to get unified support and agreement from governments around the world.

While we disagree with the idea of pausing AI development as it is unrealistic and not practicable, the open letter does clearly highlight some of AI's potential downsides. BCS believes that placing competent and capable IT professionals at the heart of AI development teams is one way to manage the risks of harms in AI deployment.

How does this fit with BCS and plans for UK AI regulation?

BCS, along with The Royal Statistics Society, National Physics Laboratory, the Alan Turing Institute, The Operational Research Society and the Institute for Mathematics and its Applications is part of the Alliance for Data Science Professionals. This means that registration, professional standards and responsible computing will form the basis of our approach to engage with AI safely.

The government recently issued its UK AI Regulation White Paper, which highlights five principles for UK regulators to achieve responsible AI in their specific sectors.

Section 4 recognises a central role for tools in trustworthy AI, including technical standards and assurance techniques to help implement the principles. It proposes a layered approach to standards, requiring regulators to consider the relevance of technical standards, encourage their adoption by actors across the AI life cycle, and to support the integration of the principles into the design, development and use of AI.

The paper also sets out the factors that the government believes regulators may wish to consider when providing guidance on, or implementing, the five principles. These refer back to international technical standards such

[16] https://www.insiderintelligence.com/insights/artificial-intelligence-healthcare/
[17] https://www.newscientist.com/article/2330866-deepminds-protein-folding-ai-cracks-biologys-biggest-problem/
[18] https://www.forbes.com/sites/markminevich/2022/07/08/how-to-fight-climate-change-using-ai
[19] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
[20] Safety, security and robustness; Appropriate transparency and explainability; Fairness; Accountability and governance; and Contestability and redress.

as ISO/IEC – such as the 'accountability and governance' principle.
The technical and operational standards provide management systems, processes and measurement methods to support implementation of the five principles. Organisations should align themselves with these best practices and ensure that their AI technology providers also adopt them as and when they are published.

The UK stance is that a new regulator is not necessary, as existing laws are usually sufficient to manage the risks of AI. But, to further support regulators as they implement the principles, the UK is proposing to establish a set of central functions to monitor the impact of the framework and assess AI risks as the technology evolves. This will identify emerging risks and opportunities, and inform targeted interventions as required.

The white paper places a clear emphasis on collaboration, recognising that the government will need to bring together voices from across the economy and wider society in order to maximise the benefits of the growth in AI while tackling some of the risks.

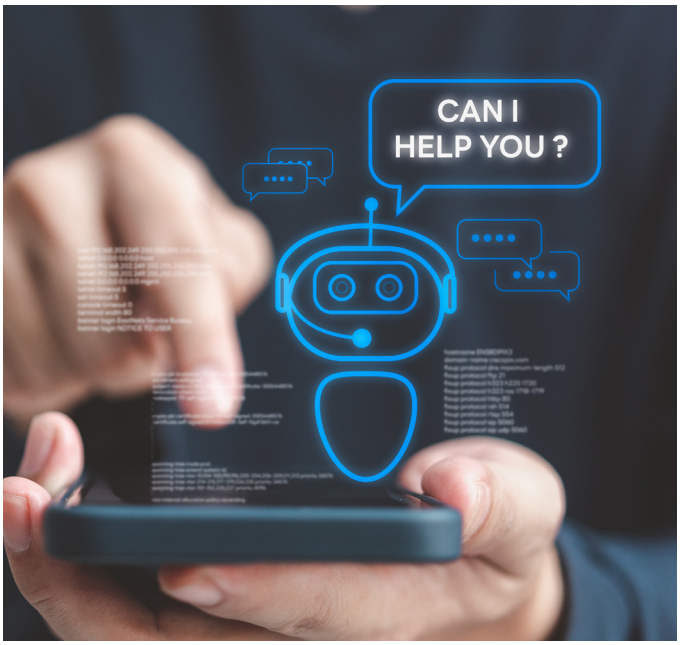
The government is currently consulting on the white paper and will continue to engage with stakeholders including industry, academia, regulators, and others to deliver a regulatory regime that is adaptable, collaborative and trustworthy.

What are the immediate options that fall short of a 'pause'?

The options between what we have now (open access via a UI/API to a black box) and a 'pause' are in the main independent audit, sandboxes and carefully monitored usage. All have different features, but all require the 'intended use' context from the person deploying the AI (not just the technology provider).

Independent audit is required by the proposed EU AI Act for high-risk AI as 'conformity assessment'. However, currently there are no independently accredited certification schemes or auditors. Typically in the UK a certification scheme is used for everything from PPE to telecoms chips – products or services are audited against certifiable standards, and the auditors are accredited by the UK Accreditation Service. This accreditation is separate from, but is supported by, professional registration. This route cannot formally progress until international technical standards are ready, and that is not likely to occur for at least another 12 to 18 months.

Areas that are already safety-critical, or covered by strict product legislation such as medical devices, are already required to undergo conformity assessment. This could be a reason why the use of AI in healthcare is so advanced.



We are aware that there are many companies offering audit-like services in other areas, but unfortunately they are no more regulated than AI development itself. Continued public use with strong health warnings, informed consent, active monitoring and the ability for the subjects of AI systems to exercise various rights are also types of assurance tools that can be used.

[21] https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/
[22] https://www.ucas.com/data-and-analysis/undergraduate-statistics-and-reports/ucas-undergraduate-releases/ucas-undergraduate-applicant-releases-2023-cycle/2023-cycle-applicant-figures-25-january-deadline

We also believe that the updated guidance issued by the Information Commissioner on the 15th of March 2023 in relation to AI is also useful for organisations to consider, as it covers the areas of accountability and governance implications of using AI; transparency; lawfulness; accuracy and statistical accuracy; fairness (including bias and discrimination and the impact of Article 22 UK GDPR on fairness); and individual rights. The ICO says the updated guidance is in line with the government's ambitions to adopt a pro-innovation approach to AI with embedded principles of fairness. Applications for computing degrees are growing faster than for any other degree subject according to the latest UCAS data; but we need digital skills for young people and adults to be far more widespread and build general understanding of AI and its relevance, opportunities and risks in everyday life. Continued support for a relevant and well-funded computing curriculum in schools is also critical.

Finally, sandboxes have been shown to be an effective way to drive innovation and regulatory learning quickly. In fact, it has been shown in the UK to be 40% faster at getting products to market, compared to the normal regulatory timeframes . They are a safe space for regulators to monitor AI systems and work with the providers, creating better products and better regulatory guidance.

We recommend that all of these approaches are explored and potentially combined to fast-track the creation of an effective AI assurance ecosystem as envisaged by the CDEI.

About BCS, The Chartered Institute for IT

BCS, The Chartered Institute for IT is the professional body for information technology. Our purpose, as defined by Royal Charter, is to promote and advance the education and practice of computing for the benefit of the public. With over 67,000 members, BCS brings together academics, practitioners, industry and government to share knowledge, promote new thinking, inform the design of new curricula, and shape policy.

About BCS' Fellows Technical Advisory Group (F-TAG)

BCS' Fellows Technical Advisory Group (F-TAG) aims to explain the opportunities and challenges of emerging technologies, what they mean for the digital industries and for the competitiveness of 'UK Plc'.

F-TAG is chaired by Adam Leon Smith CITP FBCS, and is made up of over 20 BCS Fellows selected from diverse demographic and professional backgrounds; their expertise ranges from software testing to Smart Homes.

[23] Jon Truby, Rafael Dean Brown, Imad Antoine Ibrahim, and Oriol Caudevilla Parellada. A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications. European Journal of Risk Regulation, 13(2):270–294, June 2022. Publisher: Cambridge University Press. URL: https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/ sandbox-approach-to-regulating-highrisk-artificial-intelligence-applications/C350EADFB379465E7F4A95B973A4977D, doi:10.1017/err.2021.52.

[24] The roadmap to an effective AI assurance ecosystem, Centre for Data Ethics and Innovation, 2021.

[25] As part of the National AI Strategy, the UK committed to support the National Centre for Computing Education to ensure programmes for children in AI are accessible and reach the widest demographic.