Written evidence submitted by The Department of Computing, Goldsmiths, University of London

Response to call for evidence

Communications and Digital Committee

Inquiry: Large language models

Submitted by:
Dr. Dan McQuillan, Lecturer in Creative and Social Computing

1st September 2023

The following submission represents the collated views of an expert in the social impacts of AI from the Department of Computing at Goldsmiths, University of London. Dr. McQuillan is a Co-Investigator in the ESRC-funded Centre for Sociodigital Futures and author of 'Resisting AI' (Bristol University Press, 2022).

Executive summary

1. Large language models contain foundational flaws which mean they are unable to live up to the hype and make it likely that the current bubble will burst. They will continue to require vast amounts of invisibilised labour to produce, but will not result in any form of artificial general intelligence (AGI).

2. The greatest risk is that large language models act as a form of 'shock doctrine', where the sense of world-changing urgency that accompanies them is used to transform social systems without democratic debate.

3. The AI White Paper promotes populist narratives about AI adoption that align with the hype around large language models while offering a fairly thin evidence base. Ongoing developments in UK policy, such as the upcoming summit, cite notions of existential threat while ignoring the more mundane risks of social and environmental harms.

4. The narrative around open source AI is a complete red herring. The way 'open' can be applied to large language models doesn't level the playing field, make the models more secure or challenge the centralisation of control.

5. UK regulators are not well placed to address the issues raised by large language models because these systems operate across sectors and technical, economic and social registers while establishing unpredictable feedback loops between them. Meanwhile the AI industry is already engaged in significant lobbying at the EU which has proven sufficient to dissolve regulatory red lines.

6. Additional options for regulation draw on frameworks like post-normal science to mandate an extended peer community and the inclusion of previously marginalised perspectives. This more grounded approach has a better chance of resulting in AI that is more socially productive, where regulators are supported by distributed and adaptive 'councils on AI'.

Question 1: How will large language models develop over the next three years?

5. There will be interesting technical developments in large language models over the next three years, but none of these developments will overcome the foundational problems that prevent them from being trustworthy, unbiased or truly productive. In large language models, the most intractable flaw is that their operations are optimised on plausibility not causality. In other words, they generate responses which are statistically similar to those in their training data set, refined by a set of additional guidelines for believability and non-toxicity but with no mechanism for checking facticity, so we will never be able to fully believe them even when they 'sound right'. The post-hoc methods used to constrain their 'hallucinations' (fabrications), such as Reinforcement Learning from Human Feedback[1] [2], also contribute to the way their answers to questions will change over time. In addition, the pairing of vast training datasets with the production of large quantities of misinformation looks set to progressively degrade the systems as they start to consume their own output[3] [4], while there seem to be innate security flaws in the form of adversarial attacks[5] .

6. Given these foundational problems, the consequent inability of large language models to live up to the hype, and the significant unanswered questions about the financial viability of the business models[6] [7], it is quite possible that the large language model bubble will burst within the next three years.

7. We can be certain, though, that any intermediary developments will continue to require vast amounts of invisibilised human labour both in production and ongoing operations[8]. The advent of large language models has seen a shift from simple data labelling to question and answer responses, but the scale and nature of the required labour means that a significant proportion is likely to continue as colonialised and exploitative outsourcing to the Global South[9].

8. Whatever the uncertainty about large language models over the next three years, we can say for sure that we won't see artificial general intelligence (AGI) as a result, where AGI is a form of AI that has comparable cognitive abilities as a human being and is able to apply knowledge to solve complex problems in unfamiliar circumstances. This is important because implicit or explicit beliefs in emergent artificial general intelligence are a major driver of large language models, glamourising their potential and obscuring their failings.

---

[1] Training language models to follow instructions with human feedback https://arxiv.org/abs/2203.02155
[2] Illustrating Reinforcement Learning from Human Feedback (RLHF) https://huggingface.co/blog/rlhf
[3] AI-Generated Data Can Poison Future AI Models https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/
[4] The Curse of Recursion: Training on Generated Data Makes Models Forget https://arxiv.org/abs/2305.17493v2
[5] Universal and Transferable Adversarial Attacks on Aligned Language Models https://arxiv.org/abs/2307.15043
[6] What exactly are the economics of AI? https://garymarcus.substack.com/p/what-exactly-are-the-economics-of
[7] The Inference Cost Of Search Disruption – Large Language Model Cost Analysis https://www.semianalysis.com/p/the-inference-cost-of-search-disruption
[8] AI Is a Lot of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging — and not going anywhere https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots
[9] OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic https://time.com/6247678/openai-chatgpt-kenya-workers/

a) Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories?

9. One certainty in forecasting is that the hype around large language models and generative AI will continue. This operates in two registers; that 'large language models are as transformative as electricity'[10] or that 'large language models will fully or partially replace X% of all jobs'. The second form is apparently, but not actually, more grounded than the first, given the way the capability of large language models is grossly overstated by obfuscating the complexity of real world tasks. Estimates of their impact on roles is based on reductive comparisons to the kind of abstracted task descriptions one finds in job adverts, rather than any appreciation for the kinds of implicit knowledges, interpersonal negotiations, understanding of work cultures and basic common sense that are needed in even the most mundane jobs[11].

10. It is important to note that warning of the existential risks posed by contemporary AI is itself a form of hype, inasmuch as it sustains the belief that these are indeed powerful and globally transformative. This structural dynamic is personified by AI 'Godfather' Geoffrey Hinton's turn from promoting AI as an automated substitute for radiologists[12] to warning of harmful superintelligence[13].

11. However, there are many avenues for improving confidence in future trajectories. Some of these, as detailed further below, have a long pedigree of addressing technical questions where there is high uncertainty around facts alongside urgent social and ethical questions. All require the participation and 'expertise by experience' of those most affected by the widespread adoption of the technologies.

Question 2: What are the greatest opportunities and risks over the next three years?

12. The greatest risk posed by large language models is seeing them as a way to solve underlying structural problems in the economy and in key functions of the state such as welfare, education and healthcare. The misrepresentation of these technologies means it's tempting for businesses to believe they can recover short-term profitability by substituting workers with large language models, and for institutions to adopt them as a way to save public services from ongoing austerity and rising demand. Given the limitations that have already been referred to, there is little doubt that these efforts will fail. The open question is how much of our existing systems will have been displaced by large language models by the time this becomes clear, and what the longer term consequences of that will be.

13. There are profound implications for employment; jobs will continue to be lost, while many others will be made more precarious as workers are employed in lower status roles to fix the problems created by shoddy AI emulations. While large language models are touted as being able to pass basic medical exams, they are unable to reproduce the embodied

[10] Policy paper: A pro-innovation approach to AI regulation https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[11] GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models https://arxiv.org/abs/2303.10130

[12] Deep Learning Is Hitting a Wall https://nautil.us/deep-learning-is-hitting-a-wall-238440/

[13] AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google https://www.bbc.co.uk/news/world-us-canada-65452940

understanding that is vital to real world healthcare[14]. In the education sector large language models are seen as a way to scale personalisation learning while being unable to reproduce the relationality that underpins effective learning in the first place[15]. Meanwhile, the discriminatory and illegal consequences of applying machine learning to welfare systems are a warning for what will happen when that is amplified by the pervasive application of large language models[16] [17] [18].

14. The common dynamics here are extractivism and a transfer of control. Large language models are optimised by training on vast quantities of data extracted from the activities they are intended to emulate; resulting in, for example, artists paying to use AI tools which have been trained on the free and unlicensed capture of their own prior artistic output. The control of services that result from this is ultimately in the hands of the small number of corporations who are able to carry out operations at the necessary scale and whose revenue will then come from charging rents for those services, whether directly or through intermediaries who will build on top of the basic models.

15. The net effect is the acceleration of precaritisation, outsourcing and privatisation under the cover of over-hyped technology. This constitutes a form of 'shock doctrine'[19], where the sense of urgency generated by an allegedly world-transforming technology, one which will drive the future economy and deliver health and education solutions but might also put humanity's existence at risk, is used as an opportunity to transform social systems without democratic debate.

a) How should we think about risk in this context?

16. The concept of 'risk' formulates the problem as something that can a) be quantitatively ordered b) responded to in terms of a fixed distribution. It is unlikely that the challenges raised by large language models can be approached in this way. The highly recursive character of these systems and their entanglements with social factors mean that reductive quantification is misleading, while the idea of a risk distribution diverts attention from the structural nature of the underlying problems.

17. Going beyond a purely risk-based approach makes space for a more constructive framing. This would ask how generative AI could be reconstituted socially useful production; that is, as systems that are primarily designed to fulfill social needs, that promote health,

---

[14] I'm an ER doctor: Here's what I found when I asked ChatGPT to diagnose my patients https://inflecthealth.medium.com/im-an-er-doctor-here-s-what-i-found-when-i-asked-chatgpt-to-diagnose-my-patients-7829c375a9da

[15] Degenerative AI in education https://codeactsineducation.wordpress.com/2023/06/30/degenerative-ai-in-education/

[16] Victims now know they were right about robodebt all along. Let the royal commission change the way we talk about welfare https://theconversation.com/victims-now-know-they-were-right-about-robodebt-all-along-let-the-royal-commission-change-the-way-we-talk-about-welfare-209216

[17] "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case https://www.cambridge.org/core/journals/data-and-policy/article/hey-syri-tell-me-about-algorithmic-accountability-lessons-from-a-landmark-case/22A3086554B0486BB4BBAF2D5A33A3D0

[18] Exclusion by design: intersections of social, digital and data exclusion https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1606266?journalCode=rics20

[19] https://naomiklein.org/the-shock-doctrine/

welfare and well being, that enhance jobs and working conditions, and that are socially and environmentally sustainable.

Question 3: How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?

18. The AI White Paper doesn't address the specific capabilities of large language models as such but does repeat many of the dominant narratives about AI delivering incredible benefits. The main thrust is the need to "act quickly to remove existing barriers" to AI adoption and to avoid "unnecessary regulatory burdens", on the basis that this will drive growth and lead to the UK becoming an AI superpower. The evidence base offered in the paper for these claims seems rather thin. For example, the empirical data cited to support a growth in jobs referring to the AI industry itself rather than the impact on the employment as a whole. The paper draws a shaky line between prosaic services like Netflix and the notion that "AI will transform all areas of life" by "delivering wide societal benefits". Where specific application areas are mentioned they most often refer to the self-reported achievements of Google DeepMind.

19. It is notable that the only mention of environmental impacts is the claim (based on a report from Boston Consulting) that AI is "already… mitigating climate change". The climate implications of large language models are an important and unresolved factor in assessing their overall implications and deserve a more critical treatment. While much industry data is private it is clear that the scale of large language models amplifies AI's direct environmental and social effects in terms of energy use[20], emissions[21] and water consumption[22] and may have wider knock-on effects.

20. Ongoing developments in UK AI policy prioritise the fake dangers of rogue AGI while marginalising the risk that large language models and generative AI will fail to deliver economic or social value[23]. Government plans for an AI summit in late 2023 are notable as a swing from the laissez faire of the White Paper to an embrace of misleading narratives about AI's promises and threats[24]. The fact that the summit seems largely restricted to big companies and governments stands in strong contrast to the forms of participatory regulation proposed below.

21. None of the proposed measures in the AI White Paper would interrupt an AI shock doctrine, where the hype around large language models and generative AI acts as cover for an unexamined transfer of influence and control to a relatively small number of industry actors.

---

[20] Why AI is so power-hungry https://arstechnica.com/science/2020/12/why-ai-is-so-power-hungry/
[21] Energy and Policy Considerations for Deep Learning in NLP https://arxiv.org/abs/1906.02243
[22] Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models https://arxiv.org/abs/2304.03271
[23] personal participation in recent policy development workshops
[24] UK to host first global summit on Artificial Intelligence https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence

a) What are the implications of open-source models proliferating?

22. The role of open-source is a red herring as far as large language models are concerned. Inescapably huge material and technical resources are needed to develop and deploy these models. The fact that they models may be adaptable or extensible by third parties doesn't change this centralisation of control, and the duopoly of software frameworks in which these systems are developed makes it easy for companies like Google and Meta to absorb any innovations back into their closed corporate infrastructures[25].

23. The ways 'open' is used in relation to AI doesn't achieve any of the things that open-source originally laid claim to; it doesn't democratise or level the playing field, it doesn't make the systems more secure, and rather than diffusing the concentration of power it may actually deepen it[26]. Moreover, none of the proposed 'openness' addresses the core problems of opacity and brittleness (such as the tendency to hallucinate): these are inherent to the algorithms themselves and are not solvable by 'more eyeballs'. What the narrative of open-source is achieving, if anything, is to create a cloud of confusion around regulation, a sense of confusion which is currently being leveraged by large actors in the hope of exemptions from the EU's AI Act[27] [28].

Question 4: Do the UK's regulators have sufficient expertise and resources to respond to large language models? If not, what should be done to address this?

24. It is clear that the issues raised by large language models and generative AI exceed the current expertise and resources of UK regulators, because these systems are complex disruptions which operate on technical, economic, social and political registers at the same time and establish unpredictable feedback loops between them. Most forms of expertise are not constituted to deal with these kinds of entanglements[29]. Meanwhile, calls to regulate large language models in the same way as nuclear power are part of the fear-based hype around 'AI Safety'[30].

25. Experience with social media platforms suggests that if the AI industry achieves sufficient centralisation and scale it will be able to accommodate any level of financial penalty imposed by regulators[31]. Meanwhile, the process of developing the EU's AI Act

---

[25] How Open Source Machine Learning Software Shapes AI https://dl.acm.org/doi/10.1145/3514094.3534167
[26] Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807
[27] BSA Leads Joint Industry Statement on the EU Artificial Intelligence Act and High-Risk Obligations for General Purpose AI https://www.bsa.org/news-events/news/bsa-leads-joint-industry-statement-on-the-eu-artificial-intelligence-act-and-high-risk-obligations-for-general-purpose-ai
[28] Google: Submission to European Commission on the EU AI Act https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en
[29] Predatory Formations Dressed in Wall Street Suits and Algorithmic Math https://journals.sagepub.com/doi/10.1177/0971721816682783
[30] OpenAI sees the IAEA as the future model for regulating AI https://qz.com/openai-sees-the-iaea-as-the-future-model-for-regulating-1850463478
[31] Meta to face record EU privacy fine over Facebook data transfer to US https://www.reuters.com/technology/meta-face-record-eu-privacy-fine-over-facebook-data-transfer-us-2023-05-17/

shows that corporate lobbying power is sufficient to dissolve red lines and to tip matters in favour of self-regulation[32] [33]. But the underlying challenge for regulators is that large language models are complex sociotechnical systems; that is, they are entanglements of advanced technologies and social structures that require a complex sociotechnical response. The unwanted impacts of large language models and generative AI are not amenable to a simple regulatory lever any more than the innate biases of the algorithms and datasets are fixable by a simple technical correction.

Question 5: What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?

26. One methodology for dealing with urgent sociotechnical questions is post-normal science[34]. This was proposed in the 1990s as a way of positioning science within the wider matrix of social factors, especially when "facts are uncertain, values in dispute, stakes high and decisions urgent". Post-normal science recognises the limits of orthodox expertise when dealing with phenomena that are novel and complex and whose impacts extend in scale, time and severity, and it addresses this through the extended peer community; that is, an extension of the technical peer community to include those who are directly affected, alongside complementary sources of sociological, cultural and anthropological insights.

27. As highlighted in the arguments for post-normal science, and even more so in the broader field of science and technology studies, the inclusion of previously unheard or marginalised perspectives can have the net effect of producing responses which are actually more objective and robust than would otherwise be the case[35]. This is especially relevant when it comes to AI, whose predictive and inferential operations generate both unreliable knowledge and epistemic injustice (where deflated levels of credibility are given to people's own experiences)[36].

28. A more grounded approach to large language models is an opportunity to be more socially productive. The paradigmatic example here is the Lucas Plan of the 1970s, where workers in a giant arms company applied their understanding and experience to prototype alternatives, from wind generators to hybrid vehicles and kidney dialysis machines[37]. Their participatory innovation generated products that were ahead of their time in terms of sustainability and social purpose.

a) How would such options work in practice and what are the barriers to implementing them?

---

[32] Big Tech accused of shady lobbying in EU Parliament https://www.politico.eu/article/big-tech-companies-face-potential-eu-lobbying-ban/
[33] The lobbying ghost in the machine: Big Tech's covert defanging of Europe's AI Act https://corporateeurope.org/en/2023/02/lobbying-ghost-machine
[34] Science for the post-normal age https://www.sciencedirect.com/science/article/abs/pii/001632879390022L
[35] Rethinking Standpoint Epistemology: What Is "Strong Objectivity"? https://www.jstor.org/stable/23739232
[36] Epistemic Injustice: Power and the Ethics of Knowing https://academic.oup.com/book/32817
[37] The Lucas Plan and Socially Useful Production https://steps-centre.org/blog/new-paper-lucas-plan-socially-useful-production/

29. A sociotechnical approach to large language models is one that recognises the need for a complex and socially-grounded response. Building on post-normal science and aiming for socially productive outcomes like the Lucas Plan means starting from each particular context of production and application, and setting up structures that we might call 'councils on AI'. This are quite different to algorithmic audits, which are a post-facto exercises based on a reductive set of fixed criteria[38].

30. Workers in the technology industry will have insights about the likely effects of large language models, and there are examples of positive worker action in relation to contemporary AI[39]. This needs to be supported in terms of both capacities and rights. Likewise, those professions experiencing the immediate effects of large language models, from journalism to healthcare, will need ways to ensure their grounded expertise is part of the decision-making. One barrier here is the need for a critical technical understanding that can distinguish between actual technical potential and industry hype. Finally, communities themselves need ways to participate in shaping technological interventions that will affect their lives in myriad ways. Again, a barrier here is the capacity to come to technically-informed conclusions about social consequences.

31. UK experiments in re-establishing community input into technology strategy include the collaboration between Oxford University's Responsible Technology Institute and The Upper Norwood Library Hub[40], and there are certainly others. We can also draw on the history of Technology Networks, which were set up by the Greater London Enterprise Board in the 1980s to bring together the participatory potential of local communities and the technical knowledge of local polytechnics[41]. All potential councils on AI, from tech workers to professionals to local communities, will involve forms of critical pedagogy to overcome the barriers created by hype and misinformation[42].

32. These are the lineaments of a democratic regulatory mechanism, but one with a more complex and distributed base layer than we are currently familiar with. Where transformations based on advanced technologies like large language models have the tendency to move oversight and accountability further behind the barriers of technical and institutional opacity, these participatory networks are democratising and open to all. The recursion involved in distributed regulation is one where the flow is upwards from the adaptive complexity of the councils on AI through to the regulatory bodies that need to set society-wide guardrails.

c) How can the risk of unintended consequences be addressed?

---

[38] Assembling Accountability: Algorithmic Impact Assessment for the Public Interest https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

[39] The Making of the Tech Worker Movement https://logicmag.io/the-making-of-the-tech-worker-movement/full-text/

[40] UNLH Community LAB - Driven by AI https://www.uppernorwoodlibraryhub.org/whats-on-background/unlh-driverless-cars

[41] Technology networks: science and technology serving London's needs. Greater London Enterprise Board (1984)

[42] Rethinking Education as the Practice of Freedom: Paulo Freire and the promise of critical pedagogy https://journals.sagepub.com/doi/pdf/10.2304/pfie.2010.8.6.715

33. Unintended consequences are certain to follow from the rapid adoption of large language models, which is in effect a beta test being conducted at scale by a reckless industry on society as a whole. However, it is important to ask which consequences should be seen as unintended or unforeseeable. The rollout of large language models can be examined through cybernetician Stafford Beer's heuristic of POSIWID[43] which proposes that 'The purpose of a system is what it does', rather than that which its makers say it is. If large language models, as sociotechnical systems, lead to the transformation of economic and social structures in a way that concentrates power and marginalises people, then perhaps that should be understood as their purpose and, if so, they should be responded to in ways that reverse this dynamic.

34. A similarly worrying set of unintended consequences may flow from ideological commitments that have emerged in the cultures of Silicon Valley. While these are hard to pin down to a single label, the discourses of the US tech elites are increasingly anti-democratic in a broad sense[44] (from right wing libertarianism to accelerationism to so-called long termism[45]). We only have to look at Twitter (or 'X') at the time of writing to see what can happen when a technology platform that was misunderstood as a public good is shaped by an explicit welcome of far right narratives.

Further correspondence

I would be pleased to speak further about my response. Please contact Stefani Tasheva, Research Policy and Engagement Officer, S.Tasheva@gold.ac.uk.

[43] What is cybernetics? https://www.emerald.com/insight/content/doi/10.1108/03684920210417283/full/html
[44] Silicon Valley's Safe https://www.nytimes.com/2021/02/13/technology/slate-star-codex-rationalists.html
[45] The Dangerous Ideas of 'Longtermism' and 'Existential Risk' https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk