House of Lords Communications and Digital Committee: How will AI large language models shape the future and what is the right regulatory approach?

Submission from Careful Industries. Prepared by Rachel Coldicutt, OBE, Executive Director, and Tom McGrath, Policy Researcher. 04 September 2023.

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

2.1 We welcome this inquiry from the Communications and Digital Committee into Large Language Models. LLMs are an important part of the technology landscape, but they can be overlooked in high-level regulatory and policy conversations; this is in part because the umbrella term "AI" is often used to cover a wide range of technologies, products and services that operate in a vast range of contexts. This generalism can lead to ambiguity, and greater specificity is both necessary and useful for policy and regulatory discussions.

2.2 The risks and opportunities are numerous: this submission outlines four that would benefit from additional scrutiny: understanding the limitations and applicability of narrow AI; incentivising data quality; addressing environmental harms; the policy influence of corporate actors.

2.3 **Understanding the limitations and applicability of narrow AI:** While there is currently considerable levels of hype and speculation about the potential impacts of Artificial General Intelligence (AGI), the reality is that conscious machines are the preserve of Hollywood rather than computing labs. While the outputs of ChatGPT might seem uncannily accurate, this in part because they are an echo and amalgamation of similar sentences that humans have written before. Humans are sensemaking creatures who look for meaning and anthropomorphism, and this tendency can lead to an overextension of trust in automated systems.[1]

2.4 Computer scientist Meredith Broussard explains narrow AI — the kind powered by LLMs — as a "mathematical method for prediction", or "statistics on steroids".[2] Narrow AI can spot patterns and analyse certain kinds of data much more effectively than a human might, but the range of inputs it is able to take on and respond to are limited by a number of factors, including the availability of relevant data sets. This means that narrow AI is not truly general purpose, and should be deployed with caution where automated decisions may have significant impacts on the current or future life chances and livelihoods of a person or community.

[1] Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.
[2] Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, 2019.

2.6 In line with White House *Blueprint*, articulating these limitations would help ensure future development and deployment of LLMs proceeds inline with principles of Public and Planetary Benefit.

2.7 **Incentivising data quality:** One of the primary risks with the development of LLMs is their capacity to exacerbate biases and harms. It is now a well-established principle in machine learning that many LLMs and large-scale data vision sets do not draw upon representative data, which is a contributor to biassed outputs.[3] As David Spiegelhater says, "[W]hen we want to use data to draw broader conclusions about what is going on around us, then the quality of the data becomes paramount, and we need to be alert to the kind of systematic biases that can jeopardize the reliability of any claims."[4]

2.8 The demand for very large training data sets can mean the overall quality of data can be low and self-reinforcing. For instance, the corpus of data trawled by OpenAI's GPT-3 contains the English text of Wikipedia and WebText2, a dataset containing all websites linked from Reddit. Both of these data sets over-prioritise the input of a technically savvy, male-skewing demographic sub-sector: for instance, an analysis of Reddit demographics by technology magazine Alphr in 2021 uncovered "a picture of a younger, majority white and male audience with access to higher education" while one-third of all English-language articles on Wikipedia have been edited by one man, named Steve Pruitt.[5]

2.9 These risks are far greater for non-English languages, and so LLMs could put those who don't speak English fluently at a disadvantage or even harm. The vast majority of content on the internet is written in English.[6] This data is used to train LLMs, with firms typically employing 'multilingual models' to train models in languages other than English. However, this can lead to biassed and significantly flawed models, as these models will typically use *direct translations,* failing to account for different word uses, cultural differences, or slurs, which do not easily translate across languages or dialects. Furthermore, multilingual AI models can exacerbate physical harms, both in the UK and internationally. Poorly translated LLMs can encourage the spread and creation of misinformation. This has led to violence in Ethiopia and Sri Lanka, due to poorly designed models which neglected local languages, like Sinhala in Sri Lanka, and Tigrinya and Amharic in Ethiopia.[7]

---

[3] Vinay Uday Prabhu and Abeba Birhane, 'Large Image Datasets: A Pyrrhic Win for Computer Vision?' (arXiv, 23 July 2020), https://doi.org/10.48550/arXiv.2006.16923; Bender et al., 'Stochastic Parrots'; Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021).

[4] David Spiegelhater, *The Art of Statistics: Learning from Data* (London: Pelican, 2019)., p. 85.

[5] William Sattelberg, 'The Demographics of Reddit: Who Uses the Site?', Alphr, 6 April 2021, https://www.alphr.com/demographics-reddit/; Nadia Eghbal, *Working in Public: The Making and Maintenance of Open Source Software* (Stripe Press, 2020); Wikipedia, 'Steven Pruitt', in *Wikipedia*, 12 February 2023, https://en.wikipedia.org/w/index.php?title=Steven_Pruitt&oldid=1139013187.

[6] W3Techs, 'Usage Statistics and Market Share of Content Languages for Websites, August 2023', 2023, https://w3techs.com/technologies/overview/content_language.

[7] Deck, 'AI Moderation Is No Match for Hate Speech in Ethiopian Languages', Rest of World, 27 June 2023, https://restofworld.org/2023/ai-content-moderation-hate-speech/.

2.10 Early evangelists for data and digital technologies upheld the Value-Neutrality Thesis,[8] declaring technology to be a neutral tool. But data is neither neutral nor foolproof: "garbage in, garbage out" (or GIGO) is a well-known computing axiom that refers to what goes wrong when automated processes rely on poor-quality data, and changing social realities and attitudes mean that historic data sets and approaches to classification may be inappropriate inputs for automated contemporary decision-making.

2.11 One solution to this would be for Government to work with the UKSA and the Research Councils to (a) incentivize and actively steward the creation of higher quality and more representative data sets and (b) develop standards and social auditing tools for existing data sets, perhaps building on those set out in the addendum to the recent White House publication, "From Principles to Practice: A Technical Companion to the AI Bill of Rights".

2.12 **Minimising environmental impact:** LLMs — and AI models more broadly — require significant volumes of water, energy, and materials, such as metalloids and rare metals — in addition to land. For instance, GPUs, which are in incredibly high-demand during the current 'AI boom', are manufactured using an array of various rare metals, such as tantalum and palladium.[9]

2.13 The Government considers environmental harms to be 'out of scope' of the AI White Paper. The Government's approach is inadequate for addressing these harms, prioritising current growth — and pressure from international competitors — over current and long-term environmental damage. If the Government wishes to see AI models — including but not limited to LLMs — built into the future of the UK economy, we run the risk of baking dirty models into the system, harming the path to net zero.

2.15 LLMs are particularly notable for their excessive water consumption, as a result of liquid cooling used in data centres — with an estimated 700,000 litres of water required to train OpenAI's GPT-3 in Microsoft's data centres. This will contribute to worsening droughts amid wider climate collapse. The UK's water infrastructure is ill-equipped to support the growth of new data centres, due in part to the lack of new reservoirs.[10] As such, data centres are already competing with people and agriculture for scarce water resources: in three boroughs in west London the grid has run out of power to support new house building due to the number of data centres being built along the M4 corridor while in Cambridge, water scarcity is one factor slowing down the house-building ambitions that are essential to turning the city into a "science capital".

---

[8] Joseph C. Pitt, '"Guns Don't Kill, People Kill"; Values in and/or Around Technologies', in *The Moral Status of Technical Artefacts*, ed. Peter Kroes and Peter-Paul Verbeek, Philosophy of Engineering and Technology (Dordrecht: Springer Netherlands, 2014), 89–101, https://doi.org/10.1007/978-94-007-7914-3_6.

[9] Andrew Wheeler, 'What Raw Materials Are Used to Make Hardware in Computing Devices?', Engineering.com, accessed 4 September 2023, https://www.engineering.com/story/what-raw-materials-are-used-to-make-hardware-in-computing-devices.

[10] Rob Hakimian, 'The Challenge of Building More Reservoirs to Ensure UK's Water Resilience', *New Civil Engineer* (blog), 1 September 2022, https://www.newcivilengineer.com/latest/the-challenge-of-building-more-reservoirs-to-ensure-uks-water-resilience-01-09-2022/.

2.17 Solutions to mitigate the water consumption of LLMs do exist, with one example being geographical load balancing.[11] This approach helps to direct requests to data centres based on current geographic conditions, balancing them to minimise water and energy consumption. However, further solutions should be explored. The UK Government should advocate for and champion greener AI models, providing the necessary financial backing to do so.

2.18 Carbon emissions can be managed through the use of appropriate data centres — for instance, using centres powered by renewable energy to reduce carbon emissions. However, as AI adoption rapidly scales, it is inevitable that many smaller businesses will turn to fossil fuel powered data centres, which may be able to provide lower-cost solutions.

2.19 Furthermore, it is critical that energy efficiency is prioritised. As with water infrastructure, the UK's energy infrastructure is not well-equipped to manage grid constraints, as noted by the chair of the Environmental Audit Committee.[12] The development and use of inefficient data centres to power LLMs will only tighten these constraints and may reduce green energy capacity for other uses.[13]

2.20 **The policy influence of corporate actors:** There is significant research knowledge and domain expertise within industry — indeed some commentators have been raising the risk of corporate capture of AI research activities for several years.[14] Tech firms' high levels of investment in lobbying activity is well known — and it has been incredibly effective in the UK. While AI businesses should be one of the consulted stakeholder groups for policy development, the intent of the recently announced AI Safety Summit appears to be shaped by corporate concerns;[15] the failure to engage civil society and the wider research community, let alone to reflect wider public interest, is undemocratic and unrepresentative.

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

2.22 This submission makes the case that risk and opportunity should not be viewed in opposition to one another. Instead, it advocates for a responsible, rights-respecting approach to risk that prioritises Public and Planetary Benefit.

2.23 Rather than simply mitigating the perceived risks of pursuing technical opportunities, a Public and Planetary Benefits approach would actively prioritise positive social and environmental outcomes.

2.24 LLMs, and other forms of automation, can and should contribute to a more equitable prosperous society for everyone. Framing "risk" and "opportunity" as opposing forces brings

[11] Pengfei Li et al., 'Towards Environmentally Equitable AI via Geographical Load Balancing', 27 June 2023, https://escholarship.org/uc/item/79c880vf.
[12] Philip Dunne MP, 'Letter from the EAC Chair to the Secretary of State for Energy Security and Net Zero', 4 May 2023, https://committees.parliament.uk/publications/39836/documents/193860/default/.
[13] Bender et al., 'Stochastic Parrots'.
[14] Madhumita Murgia, 'Risk of "Industrial Capture" Looms over AI Revolution', *Financial Times*, 23 March 2023, sec. Inside Business, https://www.ft.com/content/e9ebfb8d-428d-4802-8b27-a69314c421ce.
[15] 'UK to Host First Global Summit on Artificial Intelligence', GOV.UK, accessed 5 September 2023, https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence.

an unnecessarily binary dimension that often leads to deadlocked discussions around perceived trade offs. But in a just society, harm should not be regarded as a necessary outcome of innovation; in fact, innovation should be calibrated to produce societal benefits.

2.25 A Public and Planetary Benefits model is also important because sociotechnical change is complex: risks and opportunities can be closely interrelated, and solutions to one perceived harm may exacerbate or cause another. Addressing particular risks and opportunities in isolation from their wider impacts can lead to a "whack a mole" approach that impedes innovation and creates uncertainty for businesses.

2.26 Moreover, as the recent Science, Innovation and Technology Committee interim report on the governance of artificial intelligence points out, there is an "imperative to accelerate the speed of public policy thinking" in this area;[16] as such, it seems desirable to take a feasible, implementable approach to both policymaking and regulatory interventions.

2.27 It is also salient to note that standing up new regulatory regimes takes time: in the case of the Online Safety regime, Ofcom, which is a well-established regulator, has been building its capabilities and capacity for at least three years. To avoid a similar lacunae, it would be essential to establish a transition strategy; this would enable the development of LLMs to continue inline with the broader social contract.

2.26 In many technology policy debates, the "opportunities" created by new technologies are depicted as necessary exponential economic or other improvements that may, at some point, cause sufficient levels of public harm to require regulatory guardrails. This *ex-post* approach means the harms caused by technologies and their applications are often not regulated until they are already widespread, and so become difficult — if not impossible — to rectify. This, in turn, gives rise to complex legislation and more complex regulatory environments; such an approach is not well suited to rapidly changing technologies and markets.

2.27 For instance, the progress of the Online Safety Bill has been impeded by the desire to specify, evidence, and remedy many specific harms, including some that have emerged since the drafting of the Bill began. This approach creates the risk that the Bill will become outdated before it reaches a statutory footing, and means that harms that have yet to fully emerge will remain unaddressed.

2.28 We are also in the early days of the AI revolution, and do not yet know how the long-term impacts of automated decisions will play out. For instance, the use of data and automated-decision making to infer A-level results in 2020 had consequences at "infrastructural, social, and individual levels" both in Britain and other countries that rely on the GCE assessment.[17]

[16] Science, Innovation and Technology Committee, 'The Governance of Artificial Intelligence: Interim Report' (House of Commons, 31 August 2023).
[17] Upol Ehsan et al., 'The Algorithmic Imprint', 3 June 2022, https://doi.org/10.1145/3531146.3533186.

2.29 Ehsan et al. have termed this diffuse set of impacts the “algorithmic imprint” and they describe the long-tail impact, or “afterlife”, of these decisions on many young people, which will continue to unfold long after the algorithm is no longer in use. This lingering afterlife can make the consequences of the uses and misuses of data difficult to spot. The fact that impacts might emerge sometimes years after the fact, in very different contexts, can mean that ethical trade-offs around data and data management are frequently settled in ways that prioritise perceived short-term benefit or the needs of the most powerful stakeholders over truly fair or just outcomes.

5.5 What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?

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

5.1 If it is not possible to predetermine every specific harm, then a more flexible approach is needed — one that is angled towards prevention and incentivising responsible innovation that delivers Public and Planetary Benefit. Such an approach cannot only rely upon *ex-post* regulatory interventions; it also requires an effective Industrial Strategy and a clear political vision that sets out the role of LLMs in a modern democracy. Together these components create a functioning and sustainable regulatory system.

5.2 In October 2022, the White House published *A Blueprint AI Bill of Rights*,[18] a set of five principles that describe what good looks like across a wide-range of use cases. The Blueprint “is a guide for a society that protects all people from these threats—and uses technologies in ways that reinforce our highest values”. The areas of focus for the Blueprint are: safe and effective systems; algorithmic discrimination protections; data privacy; notice and explanation; and human alternatives, considerations, and fallback. This is not an exhaustive list of indicators — it does not, for instance, refer to the environmental impacts of AI — but it is expansive and meaningful in its reach and intent, and addresses foundational issues that will be applicable across the majority of contexts and sectors.

5.3 During the current period of regulatory uncertainty, the UK Government could indicate its vision for the role of LLMs in both industry and society. Setting out a direction of travel in this way would unlock some regulatory uncertainty and minimise the chilling effects of a new regulatory regime on decision making and technological development, in academic and corporate context. Per the White House *Blueprint*, this might indicate principles and signal priorities for policy development and public consultation.

[18] Office of Science and Technology Policy, ‘Blueprint for an AI Bill of Rights’ (Washington: The White House, 2022).