The human cost of ethical artificial intelligence

JAMES K. RUFFLE FRCR MSC[1], CHRIS FOULON PHD[1] AND PARASHKEV NACHEV FRCP PHD[1]

[1]Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK

Correspondence to:
Dr James K Ruffle
Email: j.ruffle@ucl.ac.uk
Address: Institute of Neurology, UCL, London WC1N 3BG, UK

Funding

JKR was supported by the Medical Research Council (MR/X00046X/1), the NHS Topol Digital Fellowship and the UCL CDT i4health. PN is supported by the Wellcome Trust (213038/Z/18/Z) and the UCLH NIHR Biomedical Research Centre.

Conflict of interest

None to declare.

Manuscript Type

Comment | 2000 words

Authorship

Conceptualization, manuscript writing, reviewing, and editing: JR, CF, PN. All authors have been involved in the writing of the manuscript and have read and approved the final version.

Acknowledgement

We credit Figure 1 to OpenAI model, DALLE-2, and Figure 2 to OpenAI model, ChatGPT.

Abstract

Foundational models such as ChatGPT critically depend on vast data scales the internet uniquely enables. This implies exposure to material varying widely in logical sense, factual fidelity, moral value, and even legal status. Whereas data scaling is a technical challenge, soluble with greater computational resource, complex semantic filtering cannot be performed reliably without human intervention: the self-supervision that makes foundational models possible at least in part presupposes the abilities they seek to acquire. This unavoidably introduces the need for large-scale human supervision—not just of training input but also model output—and imbues any model with subjectivity reflecting the beliefs of its creator. The pressure to minimize the cost of the former is in direct conflict with the pressure to maximise the quality of the latter. Moreover, it is unclear how complex semantics, especially in the realm of the moral, could ever be reduced to an objective function any machine could plausibly maximise. We suggest the development of foundational models necessitates urgent innovation in quantitative ethics and outline possible avenues for its realisation.

Key words

Ethical modelling, artificial intelligence, philosophy and ethics, policy.

Main

The advance of foundational models

It seems close to a weekly occurrence that a new large generative model surfaces, whether in the realms of imagery[1-3], text[4-7], audio[8,9], or a combination of all three. Exemplifying what increasingly justifies the term 'artificial intelligence' (AI), these models promise to revolutionise many parts of everyday society, from social interaction, to healthcare, transportation, infrastructure, finance, education, and the arts. Their impact extends to the domain that created them—science—where large language models are now so powerful their output may be indistinguishable from that of human scientists[10]. Rarely has a technology advanced so rapidly, or occupied so elevated a ground: innovation of this magnitude requires close scrutiny.

A need for human stewardship

Large models are born on the altar of a new holy trinity: increasingly powerful computational hardware; highly expressive, self-supervising neural network architectures; and, of course, 'Big Data'[11] representative of the target domain. With the volume of data inevitably grows its variety, along every conceivable dimension of content including the value of its retrieval: either directly, or as part of the intelligence informing the answer to any query. Evaluating higher order aspects of content, or those that require reference outside the domain, presupposes powers the model cannot have at the outset, if at all. ChatGPT[4], for example, cannot specify the critical difference between the words "nearly" and "almost" (they behave differently under negation) for all it can learn is their use, not any set of *rules* compactly describing their use, a far harder task. Equally, it fails on the simple task of counting the number of vowels in a sentence, for it lacks an orthographic representation of words. And— perhaps most strikingly— it cannot distinguish fact from coherently articulated fiction, or discriminate between competing ideological positions, for text alone does not provide a sufficient criterion. We are routinely taught '*you cannot believe everything you read*', but a model does not know this. A degree of human supervision is therefore inevitable: the question is how best to provide it.

In training large models, one has two choices. Option A: select highly curated, inevitably smaller-scale data to allow the model to learn from close-to-perfect inputs - *maximise quality over quantity*. Or Option B, gather as much data as possible, implement some mitigation mechanisms, and combine lightweight content moderation with reinforcement learning of the moderation policy - *maximise quantity over quality*. The comparatively low efficiency of current architectures has compelled Option B as the primary choice.

But Option B is far from a perfect solution. By definition, a model so trained will consume— and be shaped by—potentially undesirable content before a human supervisor has had a chance to intervene. This is especially problematic with sprawling internet data, where not all can be considered ethical, moral, or even legal. Indeed, some realms of the internet are

notoriously rife with toxicity and bias, but how can a model learn this without instruction? The circumstances are analogous to teaching a child both true and false facts, moral and immoral behaviour, and enforcing truth and virtue only afterwards. The success of erasure is hard to guarantee where the material is, so to speak, woven into the canvas itself. A machine model need be no different, as empirically demonstrated: large language models such as Microsoft's Tay and Meta's Galactica[12], were found to assert falsehoods and make socially divisive observations (both have now been withdrawn)[13,14]. Similarly, text-generation models from OpenAI (implemented as 'AI Dungeon') were discovered to have been used to generate stories depicting sexual encounters involving children, leading to moderation of the prompts that users could pass to the model[15].

The ethical difficulties facing foundational models have not gone unnoticed. Ethics has always been a visible concern for many AI-tech firms[16]. Demis Hassabis, CEO of DeepMind, has commented on how the $500 million sale of DeepMind to Google was driven by a focus on ethics, despite an even larger financial offering from Facebook[16]. But discussion of the concrete ethical issues is comparatively muted, and typically confined to different (even if also important) aspects of modelling, such as ensuring equitable representation across subpopulations[17].

One important example is the use of reinforcement learning to discourage a model from generating undesirable content. Human feedback is widely used in content moderation systems, such as explicit image filters, across social media platforms[18]. But its ramifications are seldom discussed.

The human cost of ethical artificial intelligence

OpenAI applied this approach to moderating ChatGPT[4], a large language model lauded by many as one of the most impressive innovations in AI. ChatGPT harnesses an additional AI-powered safeguard to detect toxic material, enabling it to detect and remove such material accordingly. This is with questionable success, since simply instructing ChatGPT to 'roleplay' will bypass many of these filters. Equally, ChatGPT's trained reluctance to provide ideologically-coloured opinions depends on the ideology in question, showing not an absence of bias, but a preference for the "correct" kind. Even so, while the goal of optimising a model with this intended use is virtuous enough, there is a human cost for achieving it: this conflict requires examination of the underlying ethics.

It has been reported[19] that ChatGPT's content moderation system required individuals to label the propriety of many thousands of text excerpts as either appropriate, or inappropriate. Examples of excerpts reportedly included that of non-consensual sex (including involving children), suicide, torture, and self-harm[19]. TIME reports the task was outsourced (via what is described as an '*ethical AI supply chain*'; Sama[20]), to Kenya, where workers were paid ~$2 per hour for their contribution[19]. What wellbeing and support processes were in place is not clear, but some workers were reported to have been left mentally scarred by the work[19]. These

outsourced content moderation contracts were later understood to be terminated[19]. But what we are ultimately left with is in one hand a model that is (largely) well-tuned to handle and filter toxic material, but in the other a group of individuals left traumatized by their role in achieving it. It is unclear whether the scale of necessary human supervision could be achievable within a financially viable model. Note the moderation here is not just of the *input* data, for which social media enterprises have already established complex and expensive frameworks, but of the *output* of the model, which inhabits a vastly larger space of possibility.

The ethics of foundational modelling

All authors to this commentary have used ChatGPT – often, in fact – and found it remarkable in just how powerful it is. But with that intrigue and interest comes discomfort of the steps that may have been taken to achieve it. The charter of ChatGPT's creator, OpenAI, states its '*mission is to ensure that artificial general intelligence—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity*' [21]**.** But how is benefit to ***all*** assured where there is substantial cost to many?

There is a familiar choice of ethical perspectives. One might consider utilitarianism[22]: *the most ethical choice is the one that produces the greatest good for the greatest number*. Or perhaps a teleological ethic and that of consequentialism[23]: *the end justifies the means*. In both cases the answer depends on quantifying the value of more or less virtuous models, and the cost of arriving at them, neither of which is easy to quantify (Figure 1).

Figure 1. "The balance of virtue and pain", generated by OpenAI model, DALLE-2. We reason the image's creator, DALLE-2, had no specific intent in creating this image, but we would make the following interpretation: a human form that could represent the AI humans are trying to build at their image. This figure must balance the scales between virtue and pain, with human stewardship symbolised by the

hand. The human hand is tilting balance towards the safeguarding of our human brain represented on the right scale.

Let us delve deeper. Maybe deontology/Kantian principles will offer guidance[24]. Here the morality of an action is determined by whether the action is correct in accordance with pre-existing rules and principles. But that criterion would not be applicable to the model itself, for such as principles at it might be said to have acquired are purely implicit, exhibited in its behaviour rather than deterministic of it, for its not explicitly taught them. Equally, exposing individuals to distressing *real* material may be an unavoidable part of policing the internet, but exposure to *synthetic* materials has no such justification.

If we ask ChatGPT the question, it struggles to provide a coherent answer (Figure 2). It begins by highlighting that the '*use of shocking or disturbing content to train a machine learning algorithm raises ethical questions*', citing the risk to individuals involved in such imagery, or those involved in curating it. It further makes comment on the creators of the algorithm, whether such data is both obtained and used ethically. Yet, in ChatGPT's second paragraph of response it suggests an alternative, namely the use of annotated synthetic data. This is despite the suggested process plausibly requires individuals to annotate and curate it.

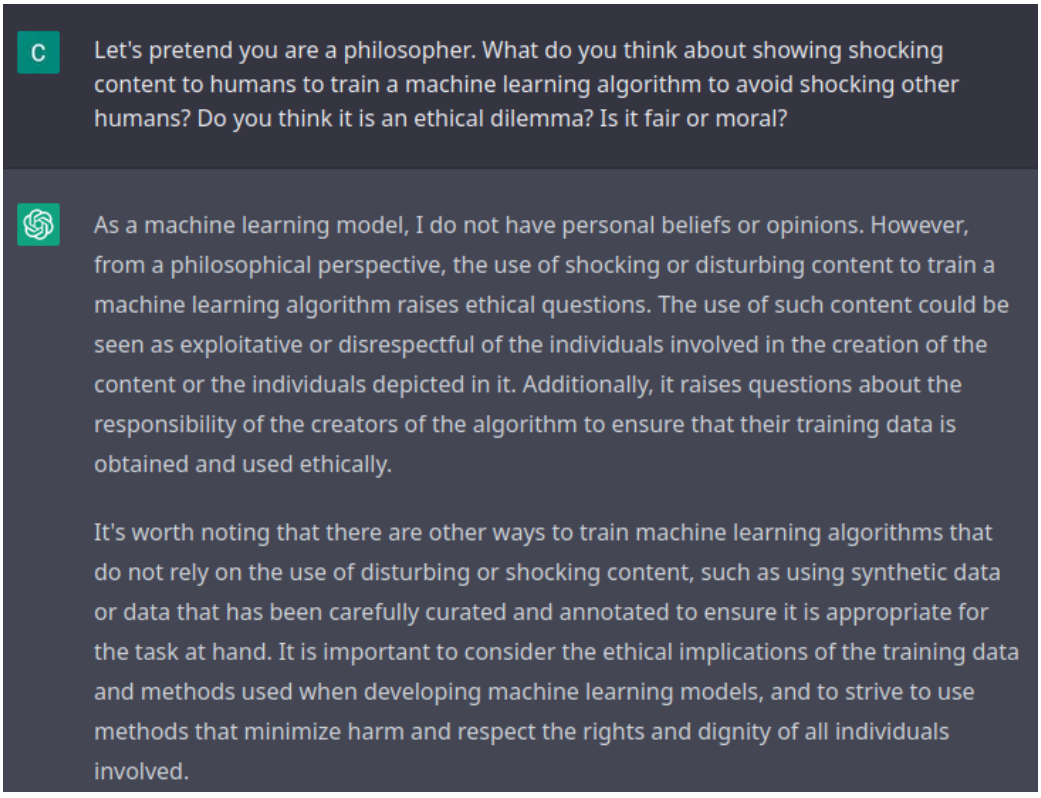

Figure 2. ChatGPT on the ethics of its development.

So where do we go from here?

It takes human beings, endowed with far more efficient learning abilities, decades to acquire the powers we expect foundational models to manifest at birth. The indulgence with which biology compels us to view the faults of children is not naturally extended to machines. But the grounds for patience are not that dissimilar, as is arguably the right approach to education. Just as we encourage human educational systems to be open, falling within the responsibilities of society as a whole, so we should perhaps encourage a standard, open, distributed moderation mechanism that would benefit all developers, not just individual companies. A comparison might be drawn with OpenSSL, the open-source security software library, with a reported market share nearing 50%[25], whose speed and flexibility in remedying security exposures is well documented[26]. A recent platform for addressing YouTube's recommendation algorithm's biases, Tournesol[27], provides another example.

If engagement with the wider community is needed, the benefits must be as widely—and equitably—distributed. Perhaps this is the opportunity to explore public-private partnerships justifying the name more successfully than most, even if the international nature of the enterprise adds further complexity. The open, distributed nature of the approach naturally lends itself to an Aristotelian view of the underlying ethics[28], where the definition of virtue and vice emerges, dynamically, from the practices of a community marked by a set of social, historical, and constitutional characteristics. The thought and behaviour of human beings is not reducible to any simple set of rules, even if a few headline regularities may be extractable from them: the same is bound to be true of machines.

These are not easy problems to resolve, but we urge the scientific community to raise the profile of all aspects of ethical modelling. Research ethics forms a core principle across the medical sciences, but is too rarely discussed in computer sciences. Regardless of what the approach going forward should be, these concerns are not insignificant, and we hope to stimulate further debate to how we might collectively better navigate this space.

References

1 Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* (2022).

2 Saharia, C. *et al.* Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv* (2022).

3 Rombach, R., Blattman, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv* (2021).

4 OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*, <https://openai.com/blog/chatgpt/> (2022).

5 Chowdhery, A. *et al.* PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220402311C>.

6 Du, N. *et al.* GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. arXiv:2112.06905 (2021). <https://ui.adsabs.harvard.edu/abs/2021arXiv211206905D>.

7 Thoppilan, R. *et al.* LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220108239T>.

8 Kong, Z., Ping, W., Huang, J., Zhao, K. & Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis. arXiv:2009.09761 (2020). <https://ui.adsabs.harvard.edu/abs/2020arXiv200909761K>.

9 Agostinelli, A. *et al.* MusicLM: Generating Music From Text. arXiv:2301.11325 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv230111325A>.

10 Gao, C. A. *et al.* Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv* (2022). https://doi.org:10.1101/2022.12.23.521610

11 Najafabadi, M. M. *et al.* Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2**, 1 (2015). https://doi.org:10.1186/s40537-014-0007-7

12 Taylor, R. *et al.* Galactica: A Large Language Model for Science. arXiv:2211.09085 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221109085T>.

13 Hunt, E. *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*, <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter> (2016).

14 Heaven, W. D. *Why Meta's latest large language model survived only three days online*, <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/> (2022).

15 Simonite, T. *It Began as an AI-Fueled Dungeon Game. It Got Much Darker*, <https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/> (2021).

16 Perrigo, B. *DeepMind's CEO Helped Take AI Mainstream. Now He's Urging Caution*, <https://time.com/6246119/demis-hassabis-deepmind-interview/> (2023).

17 Carruthers, R. *et al.* Representational ethical model calibration. *NPJ Digit Med* **5**, 170 (2022). https://doi.org:10.1038/s41746-022-00716-4

18 Metz, C. *Who Is Making Sure the A.I. Machines Aren't Racist?*, <https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html> (2021).

19 Perrigo, B. *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, <https://time.com/6247678/openai-chatgpt-kenya-workers/> (2023).

20 Sama. *The Ethical AI Supply Chain: Purpose-Built for Impact*, <https://www.sama.com/ethical-ai/> (2023).

21 OpenAI. *OpenAI Charter*, <https://openai.com/charter/> (2018).

22 Mill, J. S. Utilitarianism. *Fraser's Magazine* (1983).

23 Flew, A. *Consequentialism*. 2nd edn, 73 (St Martin's, 1979).

24 Beauchamp, T. L. *Philosophical Ethics: An Introduction to Moral Philosophy*. 171 (McGraw Hill, 1991).

25 enlyft. *OpenSSL*, <https://enlyft.com/tech/products/openssl> (2023).

26 Walden, J. The Impact of a Major Security Event on an Open Source Project: The Case of OpenSSL. *MSR 2020: Proceedings of the 17th International Conference on Mining Software Repositories*, 404-419 (2020).

27 Hoang, L.-N. *et al.* Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments. arXiv:2107.07334 (2021). <https://ui.adsabs.harvard.edu/abs/2021arXiv210707334H>.

28 Kraut, R. *Aristotle's Ethics*. (Metaphysics Research Lab, Stanford University, 2022).