Who Says?

Novel release methods of large language models and their implications for the future of artificial intelligence research

Mustafa Sadriwala

Social Issues and Ethics for Computer Science and Engineering
Dr. Douglas Dow

November 20, 2019

1 Introduction

The emergence of artificial intelligence (AI) has been one of the single largest leaps in the computation capabilities of machines in human history. It's rapid growth and ability to automate tasks has been causal in the rise and fall of entire industries and it continues to co-opt and set novel industrial practices every year. Yet, with the coming of AI technology there has – inevitably – been its adoption by malicious actors. In a recent attempt to thwart the usage of forthcoming AI research by such malicious actors, OpenAI, decided to withhold a complete release of their research and instead released only a smaller version of their model and sampling code without the underlying dataset, training code, or model weights.[1] This being the case, it becomes significantly more difficult for outside organizations and research groups to recreate or reproduce the results of their large-scale language model, Generative Pretrained Transformer (GPT-2).

OpenAI justified this choice through an analysis of the potential consequences of a complete release. They reasoned that general language models that grew to the accuracy levels of GPT-2 and beyond would be capable of potentially generating convincing fake news articles, impersonating people digitally, and automating the process of creating misleading content and phishing emails.[2] These consequences were sufficient enough for OpenAI to reconsider the publication of their entire codebase and trained models.

With AI technologies coming to the forefront of many of today's digital advances and the capabilities of AI models growing evermore, it becomes important to understand the value or lack thereof with a release pattern like GPT-2's and the underscored cautionary prescription given by OpenAI. In particular, what were the reasons behind OpenAI's decision, what are the

research community's opinions and practices surrounding it, and where does AI research go from here.

2 Dual-use Research

The topic of dual-use research has frequently been applied to the conversation of AI research, fortunately, it is not a new argument. Selgelid helps to define when dual-use research occurs: "scenarios where[in] the results of well-intentioned scientific research can be used for both good and harmful purposes give rise to ... the 'dual-use dilemma.'"[3] It's easily demonstrably that AI research and particularly the subfield of generative language models is one such scenario as several reports have been published[4,5] – in the wake of OpenAI's decision – that characterize the circumstances, opportunities, and risks of misuse.

It becomes clearer to see this analysis of AI research as being dual-use when looking at the intent of the research and seeing how those can easily be manipulated into multiple uses. For example, one report[6] highlights how an autonomous drone designed to deliver packages for large digital retailers is not very different in its capabilities than an autonomous drone meant to deliver explosives.

In analyzing the security threats of AI, there are three main domains of note: digital security, physical security, and political security. When looking directly at the effects of technologies such as GPT-2 these domains can be focused into particular areas of interest. Of note is the automation of social engineering attacks (including spam and phishing emails that are tailored to a target and made to look like one of their contacts in that contact's writing style), fake news reports, automated and personalized disinformation campaigns, and denial-of-information at-

tacks.[7]

The topic of dual-use research primarily appeared in relation to atomic physics when the study of chain reactions in unstable elements sparked the debate of potential applications in weapons and then once again rose to prominence when concerns arose over recipes for deadly viruses being studied and proliferated in the field of life sciences and biotechnology. In the former, concerns though raised by a few were ignored by some and the potentially dangerous research was published and incorporated into the subsequent development of atomic weapons by the USA in World War II.[8] The latter concern of virology research was debated by many to be a responsibility of the publishers themselves and a decision each journal must be consistent in making.[9,10] These past concerns over dual-use research demonstrate both the need for cooperation among the scientific community and coordination in publishing volatile research.

3 Release Methods

Openness has been of principal value thus far in the fast-paced AI research community and even the consideration of its reevaluation is a contentious one. Aviv Ovadya and Jess Whittlestone in *Reducing Malicious Use of Synthetic Media Research* specifically note that "focusing on the binary choice" of either publishing or not publishing "can lead the debate around openness of ML Research to become very polarized."[11] Ovadya and Whittlestone, in both *Reducing Malicious Use of Synthetic Media Research* and *The tension between openness and prudence in responsible AI research*, advocate "rethinking publication norms"[12] and outline three distinct domains in which publication can be affected: content, timing, and distribution (audience).[13,14]

Content is altered when researchers/publishers choose to not include certain elements that may allow for easier reproducibility. But content need not just be simple omission or censorship, rather the genre of AI allows for what can be thought of as a leveled censorship that involves releasing smaller datasets or smaller models (with fewer weights and capabilities) that only give general ideas about the potential of the complete model. Researchers can also decide to include harmful use case assessments (which may either usefully warn some or dangerously inspire others) and the actual theory and concepts. Timing can be useful to keep a flexible agenda: one can choose to release immediately, after a period of time, in stages, in accordance to certain events, or can make an assessment every so often as to whether to go forward with releasing or not. Distribution is tied with who to make the research visible to, it could be made available to the public at large, to anyone who upon asking is assessed for trustworthiness, to a specific set of groups labelled at a certain level of safety that this research is correlated to, or to a specific community of partners and collaborators referred to as "access communities."[15]

4 Risk Analysis & Novel Releases

In the case of GPT-2, OpenAI elected for a staged release option wherein each stage included an increase in the content (the size of the model and attached dataset) and a reevaluation of the risks in releasing that model. The releases were staged periodically and during each stage, OpenAI was afforded the opportunity to reassess and analyze the threat landscape for releasing larger models. Concurrently, they also formed partnerships with other universities and institutes doing similar research with whom they shared larger models.[16]

A large contributor in the research community, Google, has publicly shared their endeavors

to safely incorporate and govern AI research. When it comes to publishing their advances in AI, Google considers three aspects. They weigh publication benefits to the scientific community in comparison to the risk possible with their work being adopted by bad actors, the nature of the research area and whether it is unique in its advance or if likeminded research is ongoing and perceivably being published and circulated, and if there are any potential ways in which to mitigate bad actors from using the technologies for their goals – and if not whether it would be worthwhile in waiting for such potentialities.[17] It's unclear whether it is a binary decision for the company to release or not release new research but perhaps the uncertainty is caused, in part, by the argument that each new piece of research is novel and deserving of discussion about appropriate release methods.

This need for pre-publication risk assessment is echoed by the Partnership on AI (PAI) who hosted a dinner with OpenAI and other members of the AI research community following the initial release of GPT-2. Following the dinner, PAI released an article summarizing their conclusions from the evening. The attendees found several considerations worth further and more broad discussion within the community including a standardized risk assessment process to evaluate how to release novel research, a timeframe under which the review process must be held so as to not fall behind other researchers operating in the same space, a responsible disclosure process to be more transparent with delayed or withheld publications, and possibly even consideration of risks before undertaking a new scope of research. Though certain areas of priority were outlined, consensus was far from reached as attendees cohabitated multiple schools of thought including to simply remain open, conduct a review process beforehand, or only share within secure and trusted groups.[18]

Though talk of publications procedures have become largely pertinent following GPT-2's

release – or rather, lack thereof – it can be observed that past AI researchers have elected down similar, conservative paths. Of particular interest is MIRI – the Machine Intelligence Research Institute – who in November of 2018 decided to completely change their model of release. Going against the grain of a very open AI research community, MIRI, chose to switch to a "nondisclosed-by-default" methodology that essentially meant any work done within the institute would be innately internal and then only upon further consideration of the work being critical to the outside community's understanding and development would it be published. Previously, the standard was to publish any and all research whereas now nothing would be published unless it was demonstrably necessary or beneficial to publish it thus switching the onus of proof for necessity from not publishing to publishing.[19,20]

The reasoning behind risk analyses like those aforementioned are analogous to conducting a cost-benefit analysis. The 'benefit' part of this analysis lies mainly in the release of the research which may seem like it could have a fixed value across analyses but doesn't as is proven by Google's evaluation of whether a piece of research is novel in the community or whether similar research is already being conducted. Furthermore, because of the idea of leveled censorship discussed previously, a cost-benefit analysis becomes increasingly pertinent since one can find a particular size of model for which the benefits of its publication outweigh its costs. The 'cost' half of the analysis relates to the possible risks of publishing a certain model but not only just contemporary risks as one must also account for the possible uses that are not too distant or difficult to achieve from a certain model. The evaluation of future applications of a model can become difficult to understand especially in a field such as AI where innovations can happen every day.

This type of framework would account for contextual relevancy and flexibility (both of

which can be attributed to novel release methods) but also – as will be discussed later in this paper – allows plenty of room for debate since many of the factors involved in choosing a publication method don't have explicit monetary costs or markets within which to place them – as is conventional with cost-benefit analyses. This would introduce subjectivity and complications in determining what the value of the costs and benefits actually are, obfuscating the whole matter further.

5 Cooperation

Though OpenAI withheld their larger models from the public primarily, they did not shy away from sharing their data, models, and results with other researchers. Initially, they began without much infrastructure and grew their trusted circle simply based on who requested further communication about the research and were similarly aligned in their publication cautions and were deemed trustworthy.[21] Eventually, they partnered with 4 separate universities (Cornell University, The Middlebury Institute of International Studies, The University of Oregon, and The University of Texas at Austin)[22] and formed what Whittlestone and Ovadya might refer to as an access community.[23]

Even many organizations that did not directly partner with OpenAI but conducted and concluded similar AI research chose to not publish their larger models either. Rather, they joined OpenAI in their cautions and only released similarly sized models. Most of these organizations remained in contact with OpenAI, discussed their own analyses of the threat landscape for these models, and collaborated optimistically to release larger and larger language models.[24]

Some researchers from OpenAI have argued that AI research in its need for safe and re-

sponsible development is in fact a collective action problem. They contend that the AI research community will have a "race to the bottom" to find the minimum necessary safety precautions, accomplish the most efficient development, and deliver the fastest releases. In such a case, they believe there are five factors that will help the community overcome this collective action problem: "high trust between developers..., high shared gains from mutual cooperation..., limited exposure to potential losses in the event of unreciprocated cooperation..., limited gains from not reciprocating the cooperation of others..., and high shared losses from mutual defection...."[25] Unsurprisingly, these factors are linked to the four situations possible in a collective action problem (i.e. the prisoner's dilemma wherein two prisoners can both choose to cooperate and get shorter sentences, either could defect and sell the other out for the shortest possible sentence, or both could defect and get the maximum sentence). The main argument for the collective action assessment of AI research stems from the competitive pressures between organizations in the same field racing towards the same goal.[26]

However, they contend that their line of thinking is flawed since there is not always a clearly defined competitive advantage for being the first to publish research nor a definitive goal, but there can be a base assumption that sometimes in certain high-intensity fields of research with existing products and applications this type of competition does exist and can lead to a collective action problem.[27] To further cooperation in the community, the paper recommends better transparency on opportunities for collaboration, collaboration when similar research is being conducted across groups, allowing for outside, neutral parties to oversee and regulate work, and incentivizing responsibility and safety in the field through social, economic, legal, or domain-specific (e.g. greater computational power rewarded to those that abide to such standards) means.[28]

It's worth noting that though most contemporaries in the same area of research reached out to OpenAI and discussed ethical deliberations on the threat landscape of releasing variably-sized models, not all were in absolute agreement.[29] Before OpenAI released their largest GPT-2 model (containing roughly 1.5 billion parameters and nicknamed 1558M) in November of 2019, Salesforce released a 1.63 billion-parameter conditional language model (CTRL) whose generation was more controllable[30] in September of the same year.[31,32] Alongside the release of this model, Salesforce also published an analysis of the impact pretrained AI models (e.g. CTRL) can have on society and potential governance strategies for containing negative externalities.[33,34]

This same cooperation can be seen being violated in the previous anecdote of atomic physicists where the failure of some to cooperate led to the weaponization of their research. By considering the want to publish first as a collective action problem, there can be an analysis into how to minimize the defection from cooperation. The suggestion to collaborate and open projects up to collaboration nullifies the conception of competition, as organizations in the same field would join the same 'team' and no longer need to compete. Regulation makes sure that the bottom line for safety precautions doesn't fall at a level that is unacceptable for society and incentives mean that organizations won't want to ever race to this bottom line anyway, ensuring that cooperation at a sufficient level of safety for the public maintained.

6 Openness

The AI research community has long since stood on grounds of open-source[35] and collaboration, of which OpenAI has been a large proponent: open-sourcing most of their previous

projects.[36] While steps by OpenAI to more cautiously approach future research have been in-tune with conservative efforts and ideals aforementioned, many members of the community see the step as unnecessary in the face of overexaggerated risks[37,38,39,40] played up by OpenAI in what some see as a promotional stunt.[41,42,43]

Though the idea of raising concerns about safety and precaution against malicious use-cases in AI research is not inherently bad it seems poorly placed in this moment of time. OpenAI's paper is seemingly nothing out of the ordinary and is rather a natural advancement in the field of natural language processing – expected and welcomed – to suddenly promote the dangers of this area of research seems unprovoked and unnecessary.[44,45] Furthermore, by promoting this line of conservative publication methods OpenAI is changing the landscape of the research community, essentially withdrawing their backing of open-sourcing. A norm in which open-sourcing is unappealing would make the AI research community increasingly vulnerable to misinformation through the loss in reproducibility only achievable through having access to another's complete model/code.[46,47] Consequently, without an open-source promoting culture, research could not nearly move as fast as it currently does where people can build and adapt off one another's improvements without the hassle of 'rebuilding the wheel.'[48] Finally, and perhaps the most concerning result of OpenAI and others' conservative mindset is the consolidation and internalization of AI supremacy in the hands of the few computational elites that are able to afford and therefore compete in a landscape where extensive power and resources are required to create such large models.[49,50] Such a move makes this research inaccessible to smaller organizations and individuals who have previously been seen to make large and novel improvements and applications on their own (e.g. Deep TabNine).[51,52]

The basis for most of the arguments for openness in AI stem[53] from what can largely be seen

as part of the scientific ethos first proposed by Robert Merton. The general idea of openness begins with Merton's conception of communism that dictates that science is a common good and must be shared publicly so that it can grow and build off itself.[54] Furthermore, the fear and actual spread of misinformation because of the lack of openness mirrors Piotr Sztompka's argument for privatization being one of the reasons for the loss of trust in science.[55] It's clear that what OpenAI is doing would definitively be a change to the current ethos of AI, yet that does not inherently reject its feasibility nor worth.

Ovadya and Whittlestone characterize the contrary views as a result of differing beliefs on the actual versus expected risks of publication, the efficacy of novel publication methods, and how near future applications in misuse are.[56] In their paper, they propose various solutions and compromises to these disparate beliefs. When it comes to the risk of power concentrations in AI, standardized release methods that are overseen by an external organization could help in regulating misuse as well as keeping large organizations from making AI research exclusive. Arguments against the efficacy of novel release methods cannot really be proven for or against until new methods are tested empirically and are then evaluated to see which methods might be the best to utilize in the future. Finally, the quick-natured development of AI render line-drawing arguments that say current technologies are not dangerous enough to warrant ethical consideration negligent because new, increasingly mis-usable technologies are unlikely to announce themselves in their arrival or timing and might not be as far off as we expect.[57]

7 Conclusion

In consideration of further ethical discussions in the AI community it seems prudent to also look to the developers in the pivotal roles making decisions about research scopes and release methods. One study, considering accountability, responsibility, and transparency,[58] found that developers recognize all three principles of AI ethics but rarely discuss responsibility, do not actively pursue transparency, and fail to escalate internalized ethical concerns of the impact of a system.[59] The authors note that the small-scale study – though possibly under representative – highlights the disparity that exists between the theoretical ethical research conducted in AI and the reality of developers conducting research and producing the actual systems.[60]

Ovadya and Whittlestone's paper also calls for research into and eventual adoption of existing procedures in other dual-use research areas. Particularly of interest are topics of biosafety, computer/information security, and Institutional Review Boards (used in protecting human subjects of biomedical and behavioral research). Potential procedures and practices useful in the release process of AI research could include the advent of an administrative authority, specific training in release processes, formalized options for release and an accompanying rubric to decide between them, and coordination of distribution.[61]

As further debates and discussions take place as to the appropriateness of varying ideals and methodologies to reduce potential malicious use, it seems more and more likely that some form of governance will be necessary. Salesforce, though disagreeing with OpenAI's choice of self-governance, offer a wide range of analyses and recommendations to governance models in the context of AI research goals causing Holy Grail performativity.[62,63] They resist self-governance and producer-focused governance as not being inclusive enough and reject static

and dead-reckoned governance for not being flexible enough to the unpredictable landscape of AI development.[64] These holes pinpointed in current governance methods will hopefully allow for the recognition and eventual creation of an appropriate and standardized governance ideology for AI that is both adaptable enough to withstand new rule-breaking innovations and inclusive enough to avoid privatization of knowledge – which is, in fact, the unilateral goal of all the participants in the current debate over GPT-2.

Notes

[1]Alec Radford et al., "Better Language Models and Their Implications," OpenAI, February 24, 2019, https://openai.com/blog/better-language-models/

[2]ibid.

[3]Michael J Selgelid, "Governance of dual-use research: an ethical dilemma," *Bulletin of the World Health Organization* 87 (2009): pp. 720.

[4]Aviv Ovadya and Jess Whittlestone, *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, 2019, arXiv: 1907.11274 [cs.CY]

[5]Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, 2018, arXiv: 1802.07228 [cs.AI]

[6]Ibid., pp. 16.

[7]Ibid., pp. 23–29.

[8]Selgelid, "Governance of dual-use research: an ethical dilemma," pp. 720.

[9]Ibid., pp. 721–23.

[10]Thomas Ploug, "Should all medical research be published? The moral responsibility of medical journal editors," *Journal of medical ethics* 44, no. 10 (2018): 690–694

[11]Ovadya and Whittlestone, *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, pp. 7.

[12]Jess Whittlestone and Aviv Ovadya, *The tension between openness and prudence in AI research*, 2019, pp. 3, arXiv: 1910.01170 [cs.CY].

[13]Ibid.

[14]Ovadya and Whittlestone, *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, pp. 7.

[15]Ibid.

[16]Irene Solaiman et al., "Release strategies and the social impacts of language models," *arXiv preprint arXiv:1908.09203*, 2019, pp. 3.

[17]*Perspectives on Issues in AI Governance*, technical report (Google, January 22, 2019), pp. 19–20.

[18]Claire Leibowicz, Steven Adler, and Peter Eckersley, *When Is It Appropriate to Publish High-Stakes AI Research?*, Partnership on AI, April 2, 2019

[19]Nate Soares, "2018 Update: Our New Research Directions," Machine Inteligence Research Institute, November 22, 2018, https://intelligence.org/2018/11/22/2018-update-our-new-research-directions/

[20]Interestingly, this change was justified not only in its merits to decrease the potential release of research that could be appropriated for misuse, but also in the internal capacity it allowed for researchers to pursue new scopes without flinching away consciously or subconsciously from fear of its misappropriation or weaponization.

[21]Jack Clark, Miles Brundage, and Irene Solaiman, "GPT-2: 6-Month Follow-Up," OpenAI, August 20, 2019, https://openai.com/blog/gpt-2-6-month-follow-up/

[22]Solaiman et al., "Release strategies and the social impacts of language models," pp. 3.

[23]See Note 16

[24]Ibid., pp. 2.

[25]Amanda Askell, Miles Brundage, and Gillian Hadfield, *The Role of Cooperation in Responsible AI Development*, 2019, pp. 1, arXiv: 1907.04534 [cs.CY].

[26]Ibid., pp. 8–10.

[27]Ibid., pp. 3–4.

[28]Ibid., pp. 14–16.

[29]Solaiman et al., "Release strategies and the social impacts of language models," pp. 2–3.

[30]This model arguably has greater potential for misuse since users can better control the output of the model for malicious purposes.

[31]Ibid., pp. 2.

[32]Nitish Shirish Keskar et al., *CTRL: A Conditional Transformer Language Model for Controllable Generation*, 2019, arXiv: 1909.05858 [cs.CL]

[33]Solaiman et al., "Release strategies and the social impacts of language models," pp. 2.

[34]Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher, *Pretrained AI Models: Performativity, Mobility, and Change*, 2019, arXiv: 1909.03290 [cs.CY]

[35]Open Source Initiative, https://opensource.org/.

Open-source/open-sourcing refers to the release of code or software publicly (usually through websites similar to

GitHub) under a license compliant with the Open Source Definition (OSD) that allows software to be "freely used, modified, and shared."

[36]Hugh Zhang, "OpenAI: Please Open Source Your Language Model," The Gradient, February 19, 2019, https://thegradient.pub/openai-please-open-source-your-language-model/

[37]ibid.

[38]Anima Anandkumar, "An open and shut case on OpenAI," Tensorial-Professor, Anima on AI, February 18, 2019, https://anima-ai.org/2019/02/18/an-open-and-shut-case-on-openai/

[39]Delip Rao, "When OpenAI tried to build more than a Language Model," February 19, 2019, http://deliprao.com/archives/314

[40]Zachary C. Lipton, "OpenAI Trains Language Model, Mass Hysteria Ensues," Approximately Correct, February 17, 2019, http://approximatelycorrect.com/2019/02/17/openai-trains-language-model-mass-hysteria-ensues/

[41]Anandkumar, "An open and shut case on OpenAI"

[42]Rao, "When OpenAI tried to build more than a Language Model"

[43]Lipton, "OpenAI Trains Language Model, Mass Hysteria Ensues"

[44]Zhang, "OpenAI: Please Open Source Your Language Model"

[45]Lipton, "OpenAI Trains Language Model, Mass Hysteria Ensues"

[46]Rao, "When OpenAI tried to build more than a Language Model"

[47]Zhang, "OpenAI: Please Open Source Your Language Model"

[48]ibid.

[49]Anandkumar, "An open and shut case on OpenAI"

[50]Miguel Luengo-Oroz, "Solidarity should be a core ethical principle of AI," *Nature Machine Intelligence* 1, no. 11 (October 2019): 494–494

[51]Clark, Brundage, and Solaiman, "GPT-2: 6-Month Follow-Up"

[52]Jacob Jackson. https://tabnine.com/.

Deep TabNine is a code auto-completer built off the accessible, smaller models of GPT-2. It has since been updated to reflect the most recent releases of GPT-2 but serves as an example of an individual contributor creating applications solely based on the open-source offerings of larger AI research groups.

[53]pun intended

[54]Robert K. Merton, "The Normative Structure of Science," chap. 13 in *The Sociology of Science Theoretical and Empirical Investigations*, ed. Norman W. Storer (Chicago and London: The University of Chicago Press, 1973), 267–278

[55]Piotr Sztompka, "Trust in Science: Robert K. Merton's Inspirations," *Journal of Classical Sociology* 7, no. 2 (2007): 211–220, eprint: https://doi.org/10.1177/1468795X07078038

[56]Whittlestone and Ovadya, *The tension between openness and prudence in AI research*, pp. 2.

[57]Ovadya and Whittlestone, *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, pp. 8–9.

[58]Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson, *AI Ethics in Industry: A Research Framework*, 2019, pp. 5–7, arXiv: 1910.12695 [cs.CY].

[59]Ibid., pp. 8.

[60]Ibid., pp. 8–10.

[61]Ovadya and Whittlestone, *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, pp. 5–6.

[62]Holy Grail performativity describes the pursuit of some idealistic theoretical limit of computing potential that any and all researchers in a particular field are working towards achieving. Curiously, this concept seems to be in support of the notion that competition is a likelihood in AI research that OpenAI researchers were making as well and that they proposed would lead to a collective action problem. It seems that that same collective action problem is perhaps what incites the need for governance in Salesforce's analysis.

[63]Varshney, Keskar, and Socher, *Pretrained AI Models: Performativity, Mobility, and Change*, pp. 2–4.

[64]Ibid., pp. 9–10.

References

Anandkumar, Anima. "An open and shut case on OpenAI." Tensorial-Professor, Anima on AI. February 18, 2019. https://anima-ai.org/2019/02/18/an-open-and-shut-case-on-openai/.

Askell, Amanda, Miles Brundage, and Gillian Hadfield. *The Role of Cooperation in Responsible AI Development*, 2019. arXiv: 1907.04534 [cs.CY].

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, 2018. arXiv: 1802.07228 [cs.AI].

Clark, Jack, Miles Brundage, and Irene Solaiman. "GPT-2: 6-Month Follow-Up." OpenAI. August 20, 2019. https://openai.com/blog/gpt-2-6-month-follow-up/.

Fitch, Asa. "Readers Beware: AI Has Learned to Create Fake News Stories." The Wall Street Journal. Dow Jones & Company. October 14, 2019. https://www.wsj.com/articles/readers-beware-ai-has-learned-to-create-fake-news-stories-11571018640.

Hudson, Gabriel. *Behind the Data: Humans and Values*. UC Berkeley, April 2, 2019.

Keskar, Nitish Shirish, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. *CTRL: A Conditional Transformer Language Model for Controllable Generation*, 2019. arXiv: 1909.05858 [cs.CL].

Leibowicz, Claire, Steven Adler, and Peter Eckersley. *When Is It Appropriate to Publish High-Stakes AI Research?* Partnership on AI, April 2, 2019.

Lipton, Zachary C. "OpenAI Trains Language Model, Mass Hysteria Ensues." Approximately Correct. February 17, 2019. http://approximatelycorrect.com/2019/02/17/openai-trains-language-model-mass-hysteria-ensues/.

Lowe, Ryan. "OpenAI's GPT-2: the Model, the Hype, and the Controversy." Towards Data Science. Medium. February 19, 2019. https://towardsdatascience.com/openais-gpt-2-the-model-the-hype-and-the-controversy-1109f4bfd5e8.

Luengo-Oroz, Miguel. "Solidarity should be a core ethical principle of AI." *Nature Machine Intelligence* 1, no. 11 (October 2019): 494–494.

Merton, Robert K. "The Normative Structure of Science." Chap. 13 in *The Sociology of Science Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267–278. Chicago and London: The University of Chicago Press, 1973.

Ovadya, Aviv, and Jess Whittlestone. *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*, 2019. arXiv: 1907.11274 [cs.CY].

Pandya, Jayshree. "The Dual-Use Dilemma Of Artificial Intelligence." Cognitive World. Forbes. March 13, 2019. https://www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/#aa6c8016cf02.

Perspectives on Issues in AI Governance. Technical report. Google, January 22, 2019.

Ploug, Thomas. "Should all medical research be published? The moral responsibility of medical journal editors." *Journal of medical ethics* 44, no. 10 (2018): 690–694.

Radford, Alec, Jeffrey WuDario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, Ilya Sutskever, Amanda Askell, David Lansky, Danny Hernandez, and David Luan. "Better Language Models and Their Implications." OpenAI. February 24, 2019. https://openai.com/blog/better-language-models/.

Rao, Delip. "When OpenAI tried to build more than a Language Model." February 19, 2019. http://deliprao.com/archives/314.

Resnik, David B, and Kevin C Elliott. "The ethical challenges of socially responsible science." *Accountability in research* 23, no. 1 (2016): 31–46.

Selgelid, Michael J. "Governance of dual-use research: an ethical dilemma." *Bulletin of the World Health Organization* 87 (2009): 720–723.

GPT-2 from OpenAI: Better NLP Model and the Ethics Issues It Raises. Sia Data Science Lab. Sia Partners, November 1, 2019.

Simonite, Tom. "The AI Text Generator That's Too Dangerous to Make Public." Wired. Conde Nast. February 14, 2019. https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/.

Soares, Nate. "2018 Update: Our New Research Directions." Machine Inteligence Research Institute. November 22, 2018. https://intelligence.org/2018/11/22/2018-update-our-new-research-directions/.

Socher, Richard. "Introducing a Conditional Transformer Language Model for Controllable Generation." Salesforce. September 11, 2019. https://blog.einstein.ai/introducing-a-conditional-transformer-language-model-for-controllable-generation/.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. "Release strategies and the social impacts of language models." *arXiv preprint arXiv:1908.09203*, 2019.

Sztompka, Piotr. "Trust in Science: Robert K. Merton's Inspirations." *Journal of Classical Sociology* 7, no. 2 (2007): 211–220. eprint: https://doi.org/10.1177/1468795X07078038.

Thomas, Rachel. *16 Things You Can Do to Make Tech More Ethical, part 1*. fast.ai, April 22, 2019.

Trickey, Erick. *Morality in the Machines*. Harvard Law Today. Harvard Law School, June 26, 2018.

Vakkuri, Ville, Kai-Kristian Kemell, and Pekka Abrahamsson. *AI Ethics in Industry: A Research Framework*, 2019. arXiv: 1910.12695 [cs.CY].

Varshney, Lav R., Nitish Shirish Keskar, and Richard Socher. *Pretrained AI Models: Performativity, Mobility, and Change*, 2019. arXiv: 1909.03290 [cs.CY].

Vincent, James. "AI researchers debate the ethics of sharing potentially harmful programs." The Verge. February 21, 2019. https://www.theverge.com/2019/2/21/18234500/ai-ethics-debate-researchers-harmful-programs-openai.

Whittaker, Zack. "OpenAI Built a Text Generator so Good, It's Considered Too Dangerous to Release." TechCrunch. February 17, 2019. https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/.

Whittlestone, Jess, and Aviv Ovadya. *The tension between openness and prudence in AI research*, 2019. arXiv: 1910.01170 [cs.CY].

Zhang, Hugh. "OpenAI: Please Open Source Your Language Model." The Gradient. February 19, 2019. https://thegradient.pub/openai-please-open-source-your-language-model/.