Ethics and Information Technology (2023) 25:32
https://doi.org/10.1007/s10676-023-09703-z

ORIGINAL PAPER

Dual-use implications of AI text generation

Julian J. Koplin[1,2,3]

Published online: 29 May 2023

Abstract
AI researchers have developed sophisticated language models capable of generating paragraphs of 'synthetic text' on topics specified by the user. While AI text generation has legitimate benefits, it could also be misused, potentially to grave effect. For example, AI text generators could be used to automate the production of convincing fake news, or to inundate social media platforms with machine-generated disinformation. This paper argues that AI text generators should be conceptualised as a dual-use technology, outlines some relevant lessons from earlier debates on dual-use life sciences research, and calls for closer collaboration between ethicists and the machine learning community to address AI language models' dual-use implications.

Keywords Artificial intelligence · Dual-use · Ethics · Fake news · Natural language generation

Introduction

It is now possible to use AI to generate paragraphs of 'synthetic text' that are difficult to distinguish from text written by a human. The results can be striking. Open AI's GPT-2 made headlines in 2019 when, despite not having received any explicit training on these tasks, it was shown to be able to generate text in the style of hard news articles, opinion pieces, gossip magazines, academic papers, fantasy fiction, and military biographies, with it also being possible to fine-tune GPT-2 to produce more specialised styles of writing. GPT-2's successor, GPT-3, can write even more convincingly on a wider range of topics (Floridi & Chiriatti, 2020). Other AI language models, such as the Allen Institute of Artificial Intelligence's 'Grover' model, can generate news articles and opinion pieces that mimic the style of specific authors, newspapers and magazines (Zellers et al., 2019a, 2019b).[1] More sophisticated models are on the horizon. (Please note that this paper was written well before the release of ChatGPT, the release of which demonstrates how rapidly this field of AI research has progressed - and adds fresh urgency to the arguments below.)

These breakthroughs in AI text generation have the potential to both promote and undermine human wellbeing, depending on how they are (mis)used. This paper explores the dual-use implications of AI language models. Like other dual-use technologies, AI language models raise difficult questions about how best to balance the goal of promoting technological progress (and the goods this might achieve) with the goal of preventing malicious uses (and the harms that might otherwise ensue). Although the dual-use implications of AI text generation have been noted before (Solaiman et al., 2019), much of the analysis to date is focused narrowly on the use of staged release strategies to manage the relevant risks. One of this paper's key aims is to broaden these analyses. Drawing from bioethical analyses of dual-use research in the life sciences, I outline additional opportunities for intervention throughout the 'dual-use pipeline.' I close by mapping some of the pertinent ethical questions about the regulation of dual-use AI technology, and call for deeper collaboration between ethicists and the machine-learning community to resolve these dilemmas.

Crucially, text generated by these AI language models can be difficult to distinguish from legitimate articles authored by humans. Consider a recent study published in *Foreign Affairs,* which presented 500 respondents with one of four

✉ Julian J. Koplin
julian.koplin@monash.edu

1 Melbourne Law School, University of Melbourne, Melbourne, Australia

2 Biomedical Ethics Research Group, Murdoch Children's Research Institute, Melbourne, Australia

3 Monash Bioethics Centre, Monash University, Melbourne, Australia

[1] Grover can also generate metadata based on the main text of an article. Given the text of an early draft of this paper, Grover suggested the following titles: "Thinking about the ethics of AI language models," "The Complex Ethics of AI Text Generators," and, unexpectedly, "AI language models are not just about cat GIFs.".

news articles about the seizure of a North Korean ship. One of these articles was sourced from the *New York Times*, while the other three articles were generated by GPT-2. Among other questions, respondents were asked whether they found the article they had read to be credible. Surprisingly, most respondents found the 'synthetic' articles to be credible; the synthetic articles fooled between 58 and 72% of respondents. By comparison, 83% of respondents rated the original New York Times article as credible (Kreps & McCain, 2019). It is worth noting that the model used for this study—GPT-2's "medium" model with 345 million parameters—is significantly less powerful than GPT-2's "large" and "extra large" models. The gap in perceived trustworthiness would presumably shrink further if the study were repeated with newer and more powerful language models, such as the "extra large" GPT-2 model or GPT-3, which is much larger again.

Researchers from the Allen Institute of Artificial Intelligence have tested the convincingness of its own 'Grover' language model. They found that Grover can generate 'synthetic' propaganda that readers perceive as *more* reliable than human-written disinformation from known propaganda websites. One experiment paired a human-written article on the dangers of water fluoridation with a machine-written article designed to mimic the style of the *Huffington Post*; another compared human- and machine-generated articles on the supposed link between vaccines and autism. On average, respondents rated the machine-generated 'news' as more trustworthy than the human-generated propaganda (Zellers et al., 2019a, 2019b), perhaps because the AI-generated text more successfully adopted the style of reputable news publications. It is again worth noting that the field of natural language generation has advanced significantly since this study was conducted in 2019; current and future language models would likely produce even more compelling disinformation.

Dual-use problems

The term 'dual use' broadly usually refers to technologies that could be used for either good or bad purposes—especially when their use for bad purposes could have disastrous consequences (Selgelid, 2013, pp. 139–140). The term is sometimes taken to refer more specifically to two kinds of large-scale harm: harm resulting from malevolent use, and harm resulting from culpable negligence. (The possibility of unforeseeable accidents, on this account, do not render a technology dual use.) (Miller, 2018, Chap. 2.1). Dual-use dilemmas (or problems)[2] arise when it is unclear how to prevent the harms associated with misuse without also forgoing the benefits of the technology (Parliamentary Office of Science & Technology, 2009). They involve difficult questions about how to balance the risks and benefits of developing, disseminating, and controlling dangerous technologies.

A particularly stark example of a dual-use problem can be found in the history of nuclear science. Nuclear science has many beneficial uses—most obviously in energy generation, but also in medicine, agriculture, construction, and the development of electronic technologies. However, not all applications of nuclear science are benign. It also has harmful uses, most notably in the development of atomic weapons. Consider the risks associated with the current stockpiling of nuclear weapons. Even by the most cautious estimates, nuclear war risks catastrophic global harm.[3] When scientists in the first half of the twentieth century discovered atomic fission and the chain reaction, they confronted serious dilemmas about whether to publish these discoveries, given their potential to inflict massive harm (Evans, 2013; Selgelid, 2013). These discoveries were ultimately published by nuclear scientists in France, contributing to the development of nuclear weapons, the bombing of Hiroshima and Nagasaki,[4] and the ensuing nuclear arms race. Had nuclear scientists responded to the dual-use dilemma differently, it is possible that the threats posed by nuclear war could have been avoided—albeit at the expense of the positive uses to which nuclear science has been, and may yet be, turned.

Dual-use problems have also loomed large in the life sciences, which is where this paper focuses most of its attention.[5] These dilemmas often arise when the same body of knowledge could be used either to promote human wellbeing or to facilitate the development of biological weapons. A famous case involves the accidental production of an

[2] The term 'dual-use dilemma', though widely used, misleadingly implies that there are only two options to choose between—for example, whether or not to publish a paper. Usually there are more than two options; for example, it might be possible to publish a paper with various details redacted. 'Dual use problem' might be the more accurate term (Douglas 2013; Miller 2018, p. 8).

[3] Scouras (2019, p. 278) offers a relatively conservative estimate of the risks of nuclear war. He argues that the risks are often overstated; rather than threatening human extinction, nuclear war 'merely' threatens to cause "millions of deaths and unfathomable suffering" and "set civilization back centuries.".

[4] Whether the bombings of Hiroshima and Nagasaki was a good or a bad outcome of nuclear science is contentious (see e.g. Miles Jr 1985; Kimura 2013). However, it is less contentious to hold that the ensuing nuclear arms race (and threat of nuclear holocaust) was a very bad outcome indeed.

[5] There has also been some discussion of dual-use issues in the chemical industry, nuclear industry, and cyber-technology (for example, in relation to encryption and ransomware) which may bear additional lessons (Miller 2018). However, this paper focuses more narrowly on the life sciences, since this is the area of dual-use literature that is the most fully developed. The AI field also has some important parallels specific to the life sciences, such as strong norms of scientific openness (Schlagwein et al., 2017), which exacerbate dual-use risks.

extremely deadly strain of mousepox by Australian researchers in 2001. The researchers aimed to develop a tool for pest control by modifying the mousepox virus so it would render infected mice infertile. The new strain of mousepox, however, unexpectedly killed *all* mice infected with it—including mice that were resistant to ordinary mousepox or had been vaccinated against it. Although the research had beneficial uses (e.g., in pest control), it also raised bioterrorism concerns; the same techniques used to engineer vaccine-resistant mousepox could be adapted to develop vaccine-resistant smallpox for use as a biological weapon against humans (Selgelid, 2007).

The mousepox study, and others like it, has prompted much analysis of dual-use problems within bioethics and biological research regulation. For example, in the United States, concerns about dual-use life sciences research led to the creation of the National Science Advisory Board for Biosecurity (NSABB) in 2004. The NSABB is an interdisciplinary advisory committee whose responsibility have included developing guidelines for the oversight of dual-use research, providing recommendations on training scientists in biosecurity issues, and providing advice on the publication of specific experiments that raise difficult dual-use concerns (Shea, 2006). Dual-use issues in the life sciences are also discussed at the level of international institutions such as the World Health Organisation (2010) and throughout a substantial and growing body of academic literature.

In the life sciences, dual-use issues are by now relatively widely appreciated. A range of regulatory measures have been taken to mitigate dual-use risks, and there is an ongoing dialogue about how these measures can be improved (Palmer, 2020). Like the life sciences (and nuclear science before it), AI research can raise dual-use issues. However, compared to the life sciences, the dialogue around dual-use AI research is still at an early stage. The following section of this paper describes one particular area of AI research with dual-use concerns: natural language processing and the development of 'artificial text generators.'

AI language models as a dual-use technology

Like other dual-use technologies, AI language models could be used for both beneficial or malevolent purposes. Among other beneficial purposes, OpenAI envisages that its language models could be used to summarize text, translate text, and develop better speech recognition systems (Radford et al., 2019a, 2019b). The technology could support the development of AI writing assistants which could help draft or edit text for legitimate purposes. Early applications of GPT-3 include augmented writing tools designed to help write ad copy, job descriptions, and product descriptions (Dale, 2021). Some researchers are exploring the use of AI language models to help draft patent claims (Lee & Hsiang, 2019); others are investigating its use to help draft email responses (Thiergart et al., 2021). AI language models could potentially assist in drafting legitimate news articles; indeed, some traditional news outlets already use older AI technologies to draft certain kinds of content (Parkinson, 2019). Finally, AI text generation could be used for creative or artistic purposes. GPT-2 has already been employed in various kinds of creative projects, including novel-writing (Samuel, 2019), video game design (Robitzski, 2019), and surreal 'interviews' with people that don't exist (Koplin, 2019).

The threats posed by AI language models are more subtle than the possibility of nuclear or biological warfare. They are nonetheless far from trivial. The first and most widely discussed threat is that AI text generators could be used to generate 'synthetic' fake news (which would constitute a malevolent use of the technology). Fake news is currently written by humans and therefore requires non-trivial resources to generate at scale. AI text generators could automate the process, enabling malicious actors to easily and efficiently generate fake articles to support specific viewpoints, discredit particular political regimes, or praise or slander particular products, people, or companies.

As described above, the extant research suggests that many readers will find AI-generated content trustworthy. Interestingly, some of the existing research—including the *Foreign Affairs* study described above—used some of the least sophisticated versions of this technology. In the case of the *Foreign Affairs* study, the researchers generated their fake news stories using the 345M parameter version of GPT-2 (Kreps & McCain, 2019). This model is less powerful than GPT-2's subsequently released 762M parameter version and the 1.5B parameter version, let alone GPT-3. Similarly, the AI text generation tools used in the *Allen Institute of Artificial Intelligence* study in 2019 have been well and truly superseded by newer tools. If even the comparatively primitive versions of older AI language models can generate convincing text, then current and future models will pose an even greater threat.

AI-generated fake news poses a serious threat to civil society and democratic political institutions. Fake news—including false political news—appears to spread further and faster than the truth; people are more likely to retweet false claims than true ones (Vosoughi et al., 2018). Online disinformation can distort important forms of decision-making, including voting decisions and investment decisions. It might also indirectly contribute to increasing political cynicism and encourage political extremism (Lazer et al., 2018). If AI language models increase the proliferation of fake news, then the social, political, and economic impacts might be profound.

The second threat is that the technology can be used for astroturfing and related purposes. AI text generates could be used to automatically generate coherent, varied, and human-sounding snippets of text on particular themes or in support of certain goals. They could, for example, be used to efficiently generate social media posts that endorse a particular political view ('astroturfing') or to post a flood of negative or positive reviews for a company's products (Hovy, 2016; Radford et al., 2019a, 2019b). The ability to automatically generate synthetic text might enable malicious actors to give the false impression that a particular view is widely believed, or that a particular product is universally liked or disliked. The use of these tools might create so much online 'noise' that genuine comment is drowned out and meaningful dialogue rendered impossible.

A third threat is more subtle. The two threats described above would result from GPT-2 being maliciously. However, even bracketing out concerns about malicious use, there is a risk that if these technologies are used carelessly, they could result in a flood of low-quality content. Floridi and Chiriatti explain this possibility as follows:

> Given the business models of many online companies, clickbait of all kinds will be boosted by tools like GPT-3, which can produce excellent prose cheaply, quickly, purposefully, and in ways that can be automatically targeted to the reader. GPT-3 will be another weapon in the competition for users' attention. (Floridi & Chiriatti, 2020)

Low-quality content might be morally problematic even if it is not explicitly designed to advance a particular political agenda. One major problem here is that language models like GPT-3 are not designed to ensure its utterances are grounded in reality. Language models merely learn correlations between words in existing text; they do not learn how to make claims that are grounded in a common-sense understanding of how the world works, let alone complicated political, social, or scientific developments. It is common for GPT-3 to make basic mistakes in reasoning and wholly inaccurate claims. A recent investigation of GPT-3's reasoning and comprehension abilities found it made obvious errors in physical reasoning, social reasoning, and biological reasoning; in one striking example, text generated by GPT-3 warned that mixing cranberry juice with grape juice creates a potent poison. The authors conclude that GPT-3 is best characterised as a "fluent spouter of bullshit" rather than a "reliable interpreter of the world" (Marcus & Davis, 2020). Another investigation into GPT-3's ability to give medical advice found that GPT-3 would sometimes give accurate advice, but would also frequently make glaring errors. Here, one of the most striking failures occurred when GPT-3 encouraged a user asking for support for depression to kill themselves (Rousseau et al., 2020).[6] This points to a risk associated with the careless (rather than malicious) use of AI text generators: that they will successfully capture public attention while distorting our understanding of the world. Such careless use would involve a kind of culpable negligence on the part of those using the technology, again fitting the definition of dual use given above.

Admittedly, these threats are not entirely new. The internet already features much (human-generated) fake news, 'astroturfing', propaganda, misinformation, and bad advice. What *is* new about AI text generators is the ease with which they might make it possible generate malicious or misleading text at scale. Given the prospect for malicious and negligent use, we ought to think through the dual-use implications of AI text generation before better models are developed and made widely available.

It might be asked whether the discussion so far strains the definition of dual use. Traditional dual-use problems in the nuclear and life sciences involve threats of terrorism and massive loss of life. The threats posed by AI text generation arguably seem tame in comparison. Is it appropriate to extend the term 'dual use' to this latter domain?

I believe that it is. While the term 'dual use' is sometimes used only with respect to a narrow range of potential harmful outcomes—for example, those tethered to the misuse of a specific range of microorganisms or chemicals—such definitions risk overlooking areas of science and technology that are equally liable to misuse and pose an equal or greater threat (Resnik 2009); what matters, morally, is the risk to human life and happiness, not whether this danger involves one of a specific set of chemicals or scientific subspecialties. Of course, if interpreted too broadly, the term 'dual use' also threatens to lose its usefulness. If applied to *any* technology with even a remote risk of misuse, it would capture relatively benign areas of science—and measures to address dual use issues might pose undue burdens on such benign research (Resnik 2009). What matters is less the specific mechanisms by which harm might result, and more the degree of harm that might result.

In the case of AI text generation, the potential harms are substantial. Taken together, the threats posed by AI text generation could seriously undermine both individuals' autonomy and our societies' democracies. AI text generators pose threats to our autonomy because machine-generated disinformation could disrupt our capacity for knowing what's true, and therefore impede our ability to make informed,

[6] It is worth pointing out that OpenAI do not endorse the use of GPT-3 to give healthcare advice.

rational decisions.[7] If advances in AI cause disinformation to proliferate across the internet, it will become harder for us to realise our goals and act according to our values. It is for a related reason that machine-generated disinformation also poses a threat to democracy. Political disinformation can disrupt our ability to make rational, informed political decisions. Similarly, being inundated with political disinformation might breed political cynicism, division, and extremism, none of which are conducive to a healthy political landscape. And an influx of AI-generated clickbait might cause us more distracted, and more cut off from reality, than ever before.

Collective responsibility and the dual-use pipeline

One key question, then, how we can manage or mitigate the risk that AI language models will be misused. While there has been limited work on this topic specifically in relation to AI language models, the broader literature on dual-use problems—particularly in relation to life sciences research, but also fields such as chemical research and the nuclear industry—provides some crucial guidance.

Importantly, the process by which these risks would be realised has multiple stages, falling at different parts of the 'dual-use pipeline.' First, research with dual-use potential needs to be conceived, conducted, and actually result in a dual-use discovery. Second, this discovery (or the dual-use technology itself) needs to fall into the hands of malevolent actors. Third, it would need to be used effectively by these malevolent actors. Because many different things would need to happen before malevolent actors can inflict serious harm using dual-use technologies, there are multiple opportunities for intervention.

Moreover, as Miller (2018) has argued, responsibility to address dual use threats can be collective. Dual use problems are not only relevant to individual researchers; the decisions of (for example) institutions, professional associations, and governments can all affect the risk that dual use discoveries will be made and that they will ultimately be misused, and each bears at least partial moral responsibility for any resulting harm. This collective moral responsibility to mitigate dual use threats is best fulfilled via what is sometimes described as a 'web of prevention' (Miller, 2018; Selgelid, 2013): an integrated suite of measures, implemented by a range of actors, aimed at reducing dual-use risks throughout the dual use pipeline.

At the top of the dual-use pipeline lie decisions about what kinds of research ought to be pursued in the first place. Some share of the responsibility for dangerous discoveries falls on individual researchers, who can (and arguably ought to) decide against conducting harmful or seriously risky research (Forge 2013). Similar claims about moral responsibility have been made in other areas of research, where philosophers (and others) have attempted to convince researchers to abstain from contributing to morally problematic research. Sparrow (2012), for example, been argued that engineers have a moral obligation not to accept funding from the military. Even here, however, responsibility does not fall *exclusively* on individual researchers. Researchers' awareness of dual-use problems (and ability to successfully respond to them) can be promoted through a range of measures implemented at the very beginning of the dual use pipeline, such as increased education of scientists on dual-use problems, and the establishment and promulgation of codes of conduct that address dual use dangers.

In addition to individuals working in the AI field (and on text generation), companies, research institutions, investors and research funders could decide not to pursue, fund, or progress with research that carries sufficiently serious dual-use risks. Businesses, government funders, and individual researchers already make decisions about what kinds of research directions should be prioritized; the dual-use potential of risky research could factor into this decision-making (Selgelid, 2013; Miller, 2018; World Health Organisation, 2010). Such measures are already sometimes used in the life sciences. For example, US Government policies have been introduced that require institutional and federal review of some categories of dual-use research (Smith III and Kamradt-Scott 2014). In addition to these forms of oversight, the US National Institute of Health—which provides much of the funding for biomedical research in the US—has developed criteria to guide funding decisions for research involving enhanced potential pandemic pathogens (a particularly serious subset of dual-use life sciences research). These criteria make funding contingent on, inter alia, the risks of misuse being outweighed by the scientific benefits, as well investigators' and institutions' ability to conduct the research securely and communicate the results responsibility (National Institutes of Health, 2017).

Toward the middle of the dual-use pipeline lie decisions about the dissemination of dual-use research and technologies. Again, a range of different actors could have a role to play: journals could implement policies addressing whether (or under what conditions) dual use research could be published; publication of potentially dangerous discovery could be brought under institutional or governmental control; or an independent authority could provide either non-binding

[7] This could be conceptualised either in terms of disinformation undermining our autonomy per se (because accurate information might be considered necessary for autonomy), or in terms of it undermining the 'instrumental value' of one's autonomy (because disinformation prevents one from exercising one's autonomy in a way that would achieve one's goals). See generally: Taylor (2010).

advice or binding decisions about the publication of dual-use discoveries (Selgelid, 2013; World Health Organisation, 2010).

Interventions at the bottom of the dual-use pipeline aim to limit the damage that dual-use discoveries can cause in the event they are misused. Importantly, the degree of harm that malevolent actors can cause will depend on the context in which they operate. For example, if bioterrorists were to synthesise and deploy a 'live' polio virus, the loss of life would depend, inter alia, on the proportion of the population that has been vaccinated against it (Selgelid, 2013, p. 145). Similarly, if malevolent actors use AI to generate fake news, the damage they could thereby cause would partly depend on how well the audience can discriminate between reputable and untrustworthy news.

Opportunities for intervention

So far, most of the ethical and political debate on AI text generators has focused almost exclusively on the *dissemination of discoveries* by individual researchers and organisations midway through the dual-use pipeline (see e.g. Ovadya & Whittlestone, 2019; Solaiman et al., 2019; Brundage et al., 2018). In particular, this debate has focused on the possibility of tailoring release practices to mitigate dual-use risks. Release practices can be tailored in multiple ways, ranging from decisions about what content to release (e.g., more or less powerful models), when to release it, and who should be given access (e.g., other researchers or the general public). Release practices played a key role in OpenAI's response to the dual-use implications of GPT-2. Increasingly powerful GPT-2 models were released in stages, while OpenAI monitored for signs of misuse with each consecutive release (Solaiman et al., 2019). At the time of writing, GPT-3 had not been made publicly available; instead, access was provided via an API, which was itself limited to a select group of beta users (Dale, 2021).

While release practices provide an important opportunity for intervention in dual-use research, they are no panacea. Their usefulness depends, in part, on whether others are likely to achieve and publish equivalent breakthroughs; failing to publish a dual-use discovery will achieve little good if the same discovery is made and published by others shortly afterward. Indeed, this is the situation that Hungarian physicist Leo Szilard found himself in upon discovering the nuclear chain reaction; while Szilard debated the ethics of publishing his dual-use findings, a similar discovery was made and published in France, rendering the issue moot (Selgelid, 2013).

OpenAI's attempts to control access to their GPT-2 and GPT-3 language models face similar obstacles. At the time of writing the original draft of this paper, OpenAI's largest GPT-2 model had not been publicly released. However, this strategy proved less than fully effective. Two Masters' students from Brown University have already replicated OpenAI's largest model and released this replication online (Gokaslan & Cohen, 2019), essentially taking release decisions out of OpenAI's hands. More recently, OpenAI have released GPT-3. At present, access is restricted to certain categories of users (such as those in academia). Again, however, this strategy has been undermined by the release of alternatives to GPT-3, such as the open source GPT-J, that rival GPT-3's performance (Wang and Komatsuzaki 2021). For much the same reason, moratoria on further development of large language models would face practical difficulties ensuring that all actors (especially those beyond large AI labs) comply. It is therefore worth considering how a broader 'web of protection' could be spun, incorporating interventions both further up and further down the dual-use pipeline.

Within the AI community, some attention has also been paid to the possibility of developing technological solutions to the problem of machine-generated fake news that could be implemented at the bottom of the dual-use pipeline. Multiple teams are developing AI tools to distinguish between human- and machine-generated text, based on certain peculiarities of machine-generated language (Solaiman et al., 2019; Zellers et al., 2019a, 2019b). Such tools could theoretically be implemented by social media websites to filter out machine-generated news, and/or used by individuals to manually confirm whether a particular article has been written by a human or a machine. While these tools appeared promising for identifying text generated by GPT-2, it is unclear how well they will function for GPT-3 and other newer models.

There are other opportunities to mitigate dual-use risks that have received little discussion in the literature on AI language models. These include upstream measures aimed at preventing the most concerning types of research from taking place, as well as additional downstream measures to mitigate the harms of malicious use.

The most upstream measures might include new forms of education and training, or the development of professional codes of conduct that encourage researchers to consider risks of malevolent use before conducting potentially dangerous research. To some extent, this work has begun. Various ethical AI frameworks have already been developed and promulgated, some (but not many) of which do mention the importance of anticipating and mitigating dual-use risks (Hagendorff, 2020; Jobin et al., 2019). These initiatives, however, have been criticised for producing principles that are too vague or superficial to be meaningfully action-guiding (Hagendorff, 2020; Whittlestone et al., 2019). Perhaps more concerningly, the promulgation of ethical guidelines

does not appear to have improved ethical decision-making in AI research. Consider a recent controlled study, which presented two groups of tech professionals to a series of vignettes raising ethical issues. One group of professionals was presented with the ethics guidelines of the Association for Computing Machinery before responding; the other was not. There were no statistically significant differences in how these groups responded to any of the vignettes (McNamara et al., 2018). This is not to suggest that there is no value in incorporating dual-use issues into AI ethics frameworks. Indeed, ethics training and codes of conduct are widely seen to play an important role in biomedical research and medicine (see e.g. Stankovic and Stankovic 2014); presumably they could prove equally effective within AI research. However, it will be important to consider how these strategies can be *meaningfully* incorporated into education and research; the mere writing and promulgating of guidelines will not be enough.

Slightly further downstream, it might be worthwhile to develop a model for ethics committee review for high-risk forms of AI research. AI ethics review committees could operate similarly to existing medical research ethics review committees (see e.g. Leidner & Plachouras, 2017; Véliz, 2019). Like human research ethics committees, their scope could include dual-use risks alongside other concerns. AI ethics review committees would not need to resemble medical ethics research committees in all respects; the review process could be voluntary or compulsory, and the decisions made by review committees could be binding or non-binding, depending on how much weight we think ought to be placed on preventing the development of harmful technologies on the one hand, and facilitating technological progress on the other. Again, a range of actors have a role to play here, including research institutions such as universities, private companies, and governments (which could, inter alia, establish review processes similar to those already in place for dual-use life sciences research, discussed above). In designing these review mechanisms, it is worth keeping the parallel with life sciences research ethics in mind. On the face of it, the moral case for ethics committee oversight of life sciences research seems to apply *mutatis mutandis* to computer science research. We have good moral reason to subject potentially dangerous forms of research to review.

Still further downstream lie questions about whether to publish and disseminate (or to censor and control the dissemination of) dual-use AI discoveries. Again, many actors have a role to play here, including in the first instance the computer scientists that have made these discoveries. A robust 'web of prevention' may include measures that extend beyond individual researchers (who have self-interested reasons to disseminate discoveries that could be important to their career advancement). These potentially include the creation of boards such as NSABB to advise on the publication on dual-use discoveries, as well as more controversial measures such as governmental censorship. It should be acknowledged that governmental censorship raises its own serious issues, including the possibility that bureaucrats may fail to fully recognise (and therefore place appropriate weight on) the benefits of potentially risky AI research. Drawing on a suggestion made by Selgelid (2013, p. 150) in relation to the life sciences, the best solution may involve decision-making by a mixed panel of experts, including governmental personnel, computer scientists, security experts, and ethicists.

It might be asked whether it is too late for such midstream measures to make a difference, at least in the case of AI language models; sophisticated text generation technologies have already been developed. However, these models remain far from perfect; among other issues (and as briefly mentioned above), current models still struggle with common-sense reasoning (Zellers et al., 2019a, 2019b). For this reason, synthetic text that initially appears compelling can fail to hold up to close inspection.[8] It might be possible to mitigate the risks of AI language models by opting not to develop models that are more advanced and more convincing (and less easily detectable). Increasing the sophistication of language models might help achieve new kinds of benefits, but also increase the scope of the threats.

It is also worth considering additional downstream measures aimed at reducing the potential harm that could be wrought by AI text generators, should they be deployed for malicious purposes. As we have seen, efforts are already underway to develop tools to help identify machine-generated text. These are in essence technical solutions to the problems of AI-generated fake news. There are, however, additional downstream measures that are worth considering.

The potential damage that fake news can cause depends, in part, on how susceptible people are to fake news in the first place. Strengthening measures to combat fake news in general will help mitigate the specific threat posed machine-generated fake news. Many of the strategies being explored in relation to fake news more generally—which include potential legal interventions to penalise the distribution of fake news or encourage the spread of legitimate information (Flick 2017), adjustments to how social media platforms share 'news' with users (Lazer et al., 2018), and educational strategies to promote information literacy (Phippen et al. 2021)—would also prove useful in relation to the emerging

[8] To demonstrate the power of their GPT-2 model, OpenAI used it to generate a news article about the discovery of unicorns in the Andes Mountains. While the story mimicked the tone and style of hard news, it also included some bizarre or paradoxical claims—for example, that the unicorns were first created "when a human and a unicorn met each other" (Radford et al., 2019a, 2019b).

threats posed by AI-generated text.[9] The better equipped we are to deal with dis- and mis-information in general, the less damage AI text generators will be able to cause.

At a minimum, there is a need to meet the current dearth of detailed guidance available to AI researchers who confront dual-use issues, as well as the research institutions and/or private companies at which they work. Such guidance is more readily available in the life sciences. For example, the World Health Organisation (2010) provides a self-assessment questionnaire (alongside further advice) to highlight gaps or weaknesses in laboratories' and public health facilities' bio-risk management (including specifically in relation to dual-use issues). The development of similar resources—perhaps by an advisory body similar to the NSABB, professional organisations, or international bodies—would be useful for both AI in general and AI text generation specifically.

The above discussion has touched on many groups who are implicated in the collective responsibilities to address dual-use problems. The final section of this paper discusses the contribution that could be made by a group whose work is less obviously connected to AI research: moral philosophers.

How moral philosophy can contribute to ethical regulation of dual-use AI research

This final section of the paper maps some of the ethical issues that need to be considered when developing strategies to manage dual-use issues. It aims, in other word, to show what kind of analysis we will need to conduct to achieve ethical regulation of dual-use AI research.

Here, too, there are potentially fruitful parallels with the life sciences. Developing a strategy to manage dual-use issues requires us to weigh conflicting values. The values at stake in the regulation of dual-use AI research partly resemble those at stake in dual-use life sciences research. In both cases, the value of (scientific/technological) openness and (scientific/technological) need to be weighed against risks of harm if the technology is misused. As described above, the malicious or careless use of AI text generators threaten two important values: individual autonomy and political security. Some of the measures we could take to reduce this threat would also require us to curb openness within AI research—for example, by restricting the development or dissemination of dangerous technologies. These measures would also preclude some of the benefits that dual-use technologies might bring. In the case of AI text generators, the benefits include economic gains to companies that develop or utilise AI text generators, personal gains to those who utilise them, and cultural goods associated with artistic projects that use the technology.

As in the life sciences, openness is valued highly within information technology (Schlagwein et al., 2017). Accordingly, it might be thought that we should consider only interventions (like some of the down-stream interventions listed above) that do not violate scientific openness. However, it is implausible to think openness should *always* take priority over other values. This stance is already widely accepted in relation to bioterrorism. If scientists were to discover an easy way to produce a pathogen as deadly, contagious, and untreatable as smallpox, and if bioterrorists were highly likely to exploit this discovery to cause a massive loss of human life, then it would seem obstinate in the extreme to insist that the details for how to produce this pathogen be made public. By the same token, if certain AI discoveries would devastate our political system, then there are strong moral reasons against developing or disseminating these tools—even at some cost to openness. We have a moral imperative to prevent serious risks of catastrophic harms that can trump the moral reasons in favour of openness.[10]

This extreme example shows only that openness should not be considered inviolable. It makes sense to prioritise risk minimisation over openness when the stakes are catastrophically high. However, for many strands of dual-use research—in both AI and the life sciences—the stakes aren't weighted so starkly in one direction or the other. To address the dual-use risks in AI fairly, we will need to develop a defensible account of why and how much openness matters, as well as a framework for making trade-offs between openness and (say) the individual and political harms associated with a proliferation of fake news.

Relatedly, it might be asked whether we ought to prefer self-regulation by the AI community, external government regulation, or a combination of the two. There is an existing debate within bioethics on whether self-regulation is an acceptable approach to managing dual use risks within the life sciences. For example, it is sometimes argued that self-regulation is not effective enough to manage dual-use risks, since scientists (a) may lack the knowledge and expertise required to assess security risks, and (b) have a vested interest in promoting their own careers even if this involves

[9] One concern here is that alerting people to the possibility of AI-generated fake news might breed broader distrust of news media in general—which would have negative consequences of its own for democracy and collective decision-making. (I thank an anonymous peer reviewer for raising this point.) The aim, then, should be not to provoke global scepticism about what we read, but to limit people's exposure to fake news and increase their ability to identify it when they do encounter it.

[10] A similar point has been made in the ethics literature on dual-use life sciences research (Douglas and Savulescu 2010; Kuhlau et al., 2013; Selgelid 2007).

publishing dangerous research (Selgelid, 2007). Conversely, others argue that a well-developed system of self-regulation could strike a better balance between respecting scientific openness and protecting society from harm, at least provided that scientists engage with the system in good faith (Resnik, 2010). The same issues arise in relation to AI research (LaGrandeur, 2020). Here, too, there are difficult philosophical questions about how we ought to value the benefits of government regulation (which plausibly would more effectively protect against dual-use risks) and its drawbacks (which plausibly could impose undue costs to technological progress).

Responding to dual-use issues will also require us to make decisions under situations of risk and uncertainty. Not only is it unclear how much value we ought to attach to (say) preserving openness and preventing harm, it is often unclear whether, or even how likely it is that, the anticipated harms will occur. Under such conditions, our attitudes to risk and uncertainty could have a profound impact on the actions or policies we ultimately pursue. Our approach to managing risk is therefore itself an important ethical question (Hansson, 2010).

One influential approach is expected utility maximization, which would recommend the course of action with the highest expected utility (calculated by adding the utility of each possible outcome of that action multiplied by its likelihood). While expected utility maximization has been widely adopted, other approaches to decision-making are also worth considering. It might be the case that we should adopt some version of the precautionary principle, according to which we should be more averse to risks of serious harm than expected utility theory suggests (see generally Koplin et al., 2019). The precautionary principle is arguably especially useful for contexts where we are unable to estimate the likelihood of good and bad outcomes—which is arguably the case for much dual-use research (Gardiner, 2006). Alternatively, it might be the case that we should place extra weight on securing (rather than maximising) wellbeing—for example, by preferring options with less variance in expected utility across possible outcomes (Herington, 2016). And it might matter, morally, whether the imposition of risks on certain groups violates moral rights or raises distributive/egalitarian concerns—considerations that are not captured by aggregate utility (Hansson, 2003). Here, too, moral philosophers have a crucial role to play in untangling some of this complexity.

Conclusion

My aim here in this paper has not been to recommend any specific course of action. Instead, I have aimed, first, to outline a range of possible interventions that have not yet received much attention in relation to AI text generation, and second, tried to specify on the ethical and philosophical questions that the appropriate response to dual-use technologies (including AI text generators) turns on. This paper therefore leaves many important questions open. However, at least three key lessons can be drawn from this discussion.

First, it is important to look beyond release strategies and technological methods of identifying 'artificial text.' As the literature on dual use life sciences research demonstrates, there are a suite of opportunities to intervene in the 'dual use pipeline.' Many of these remain under-explored.

Second, it is important that we recognise that our intuitions about which kinds of research are too dangerous, or which kinds of restrictions are too heavy-handed, often turn on contestable views about the relative weight of values like security and scientific openness, as well as how we ought to manage risk and uncertainty. When GPT-2 was first announced, initial reactions ranged from intense alarm at the dangers of GPT-2, to claims that it is socially irresponsible *not* to release cutting-edge language models (and that OpenAI's tiered release strategy must therefore have been motivated by a desire to build hype around their technology) (Lowe, 2019). One central aim of this paper has been to bring the ethical questions raised by dual-use AI into sharper focus, and to encourage more careful discussion of these issues.

Third, insofar as the appropriate response to dual-use AI technology turns on difficult ethical questions, there are grounds for fruitful collaboration between ethicists and AI researchers. Much of the groundwork for this collaboration has already been laid. There have been many recent calls for AI researchers to consider the ethical facets of their work (Véliz, 2019), including specifically researchers developing AI language models (Brundage et al., 2018; Hovy & Spruit, 2016; Ovadya & Whittlestone, 2019; Solaiman et al., 2019; Whittlestone et al., 2019). In terms of dual-use issues, it is worth noting that (bio)ethicists have already elucidated many of the key issues (and possible responses to them) in relation to life sciences research (Douglas & Savulescu, 2010; Evans, 2013; Evans et al., 2015; Kuhlau et al., 2013; Miller, 2018; Resnik, 2010; Selgelid, 2013, 2016). As of yet, however, there has been little work by ethicists specifically on the dual-use potential of AI language models. This is a topic on which ethicists are well-positioned to contribute, and which could benefit from close collaboration between ethicists and the machine learning community.

10 years ago, bioethicists Thomas Douglas and Julian Savulescu argued that bioethicists needed to begin developing an "ethics of knowledge." The impetus for their argument was a series of advances in the life sciences—specifically, in the field of synthetic biology—have beneficial applications but could also be misused in biological terrorism or warfare. Given the stakes, Douglas and Savulescu argued that it had

become imperative for bioethics and the life sciences to confront ethical questions about the production, dissemination, and regulation of dangerous kinds of knowledge.

The field of artificial intelligence research today is in a similar position to the field of synthetic biology 10 years ago. We are living in a world where the line between human- and computer-generated content is becoming increasingly difficult to draw, and in which it might soon become possible to spread disinformation with greater ease, and on a greater scale, than ever before. We need to begin thinking through the dual-use implications of AI text generators and other technologies, not just in relation to release practices but at every stage of the dual-use pipeline. Fortunately, we currently have a window of opportunity in which we are aware of the dual-use risks, but the technology has not yet been turned to malicious ends (Solaiman et al., 2019). We ought to develop an ethical framework before this window closes. Moreover, any work done in this space will prove valuable for future discussions of dual use problems in AI more generally.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. JK acknowledges the support of the Victorian State Government through the Operational Infrastructure Support Program.

References

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., & Filar, B. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint. http://arxiv.org/abs/1802.07228

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113–118.

Douglas, T., & Savulescu, J. (2010). Synthetic biology and the ethics of knowledge. *Journal of Medical Ethics, 36*(11), 687–693.

Evans, N. G. (2013). Contrasting dual-use issues in biology and nuclear science. In B. Rappert & M. Selgelid (Eds.), *On the dual uses of science and ethics*. ANU Press.

Evans, N. G., Lipsitch, M., & Levinson, M. (2015). The ethics of biosafety considerations in gain-of-function research resulting in the creation of potential pandemic pathogens. *Journal of Medical Ethics, 41*(11), 901–908.

Flick, D. 2018. "Combatting Fake News: Alternatives to Limiting Social Media Misinformation and Rehabilitating Quality Journalism." *SMU Science and Technology Law Review*. *20*(2): 375–405.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.

Forge, J. (2013). Responsible Dual Use. In B. Rappert & M. Selgelid (Eds.), *On the dual uses of science and ethics*. ANU Press.

Gardiner, S. M. (2006). A core precautionary principle. *Journal of Political Philosophy, 14*(1), 33–60.

Gokaslan, A., & Cohen, V. (2019). OpenGPT-2: We replicated GPT-2 because you can too. *Medium*, August 23.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120.

Hansson, S. O. (2003). Ethical criteria of risk acceptance. *Erkenntnis, 59*(3), 291–309.

Hansson, S. O. (2010). The harmful influence of decision theory on ethics. *Ethical Theory and Moral Practice, 13*(5), 585–593.

Herington, J. (2016). Health security and risk aversion. *Bioethics, 30*(7), 479–489.

Hovy, D. (2016). The enemy in your own camp: How well can we detect statistically-generated fake reviews—An adversarial study. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (Volume 2: Short Papers).

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (Volume 2: Short Papers).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kimura, T. (2013). Applying Taurek's 'Should the numbers count?' to (un)justify Hiroshima and Nagasaki: A combination of historiography and applied ethics. *Flinders Journal of History and Politics, 29*, 20–40.

Koplin, J. (2019). The very human language of AI. *Pursuit*, September 8.

Koplin, J., Savulescu, J., & Gyngell, C. (2019). Germline gene editing and the precautionary principle. *Bioethics, 34*(1), 49–59.

Kreps, S., & McCain, M. (2019). Not your father's bots: AI is making fake news look real. *Foreign Affairs*, August 2.

Kuhlau, F., Höglund, A. T., Eriksson, S., & Evers, K. (2013). The ethics of disseminating dual-use knowledge. *Research Ethics, 9*(1), 6–19.

LaGrandeur, K. (2020). How safe is our reliance on AI, and should we regulate it? *AI and Ethics, 1*, 1–7.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., & Rothschild, D. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

Lee, J.-S., & Hsiang, J. (2019). Patent claim generation by fine-tuning OpenAI GPT-2. arXiv preprint. http://arxiv.org/abs/1907.02052

Leidner, J. L., & Plachouras, V. (2017). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL workshop on ethics in natural language processing*.

Lowe, R. (2019). OpenAI's GPT-2: The model, the hype, and the controversy. *KDnuggets*. https://www.kdnuggets.com/2019/03/openai-gpt-2-model-hype-controversy.html

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 2018.

Miles, R. E., Jr. (1985). Hiroshima: The strange myth of half a million American lives saved. *International Security, 10*(2), 121–140.
Miller, S. (2018). *Dual use science and technology, ethics and weapons of mass destruction*. Springer.
National Institutes of Health. (2017). *Framework for guiding funding decisions about proposed research involving enhanced potential pandemic pathogens*. NIH.
Ovadya, A., & Whittlestone, J. (2019). Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. arXiv preprint. http://arxiv.org/abs/1907.11274
Palmer, M. J. (2020). *Learning to deal with dual use*. American Association for the Advancement of Science.
Parkinson, H. J. (2019). AI can write just like me. Brace for the robot apocalypse. *The Guardian*, February 16. https://www.theguardian.com/commentisfree/2019/feb/15/ai-write-robot-openai-gpt2-elon-musk
Parliamentary Office of Science and Technology. (2009). *The dual-use dilemma*. Parliamentary Office of Science and Technology.
Phippen, A., Bond, E., & Buck, E. (2021). Effective strategies for information literacy education: combatting 'fake news' and empowering critical thinking. In *Future Directions in Digital Information* . (pp. 39–53). Chandos Publishing.
Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundin, M., & Sutskever, I. (2019a, February 14). Better language models and their implications. https://openai.com/blog/better-language-models/
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019b). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8).
Resnik, D. B. (2009). What is "dual use" research? A response to Miller and Selgelid. *Science and engineering ethics, 15*(1), 3–5.
Resnik, D. B. (2010). Can scientists regulate the publication of dual use research? *Studies in Ethics, Law, and Technology*. https://doi.org/10.2202/1941-6008.1124
Robitzski, D. (2019). A neural network dreams up this text adventure game as you play. *Futurism*, September 5.
Rousseau, A.-L., Baudelaire, C., & Riera, K. (2020). Doctor GPT-3: Hype or reality? *Nabla*.
Samuel, S. (2019). How I'm using AI to write my next novel. *Vox*, August 30.
Schlagwein, D., Conboy, K., Feller, J., Leimeister, J. M., & Morgan, L. (2017). *"Openness" with and without Information Technology: A framework and a brief history*. SAGE Publications.
Scouras, J. (2019). Nuclear war as a global catastrophic risk. *Journal of Benefit–cost Analysis, 10*(2), 274–295.
Selgelid, M. J. (2007). A tale of two studies: Ethics, bioterrorism, and the censorship of science. *Hastings Center Report, 37*(3), 35–43.
Selgelid, M. J. (2013). Ethics and censorship of dual-use life science research. In M. L. Gross & D. Carrick (Eds.), *Military medical ethics for the 21st century*. Ashgate.
Selgelid, M. J. (2016). Gain-of-function research: Ethical analysis. *Science and Engineering Ethics, 22*(4), 923–964.
Shea, D. A. (2006). *Oversight of dual-use biological research: The National Science Advisory Board for Biosecurity*. Congressional Research Service reports.
Smith, F. L., III., & Kamradt-Scott, A. (2014). Antipodal biosecurity? Oversight of dual use research in the United States and Australia. *Frontiers in Public Health, 2*, 142.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv preprint. http://arxiv.org/abs/1908.09203
Sparrow, R. (2012). "Just Say No" to Drones. *IEEE Technology and Society Magazine, 31*(1), 56–63.
Stankovic, B., & Stankovic, M. 2014. "Educating about biomedical research ethics." *Medicine, Health Care and Philosophy, 17*, 541–548.
Taylor, J. S. (2010). *Practical autonomy and bioethics*. Routledge.
Thiergart, J., Huber, S., & Übellacker, T. (2021). Understanding emails and drafting responses—An approach using GPT-3. arXiv preprint. http://arxiv.org/abs/2102.03062
Véliz, C. (2019). Three things digital ethics can learn from medical ethics. *Nature Electronics, 2*(8), 316–318.
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, 2019.
World Health Organization. (2010). Responsible life sciences research for global health security: *A Guidance Document,* Switzerland: WHO Press.
Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019a). HellaSwag: Can a machine really finish your sentence? arXiv preprint. http://arxiv.org/abs/1905.07830
Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019b). Defending against neural fake news. arXiv preprint. hhttp://arxiv.org/abs/1905.12616