Houming Chen

Jonathan Holland

ENGLISH 125-010

28 April 2022

Open Pre-trained Deep Learning Models are Good If and Only If We are Ready for It

1. Introduction

In the middle of the 20th century, after human civilizations had spent thousands of years exploring and applying their own intelligence, they proposed a new challenge: to build artificial intelligence (AI) (Stuart and Peter 1). Inspired by the structure of the human brain, which is composed of billions of neurons (nerve cells) connecting with each other, one approach to building AI is to mimic the human brain and build artificial neural networks (Dongare et al. 189). Deep learning, which originally meant "machine learning that attempts to learn in multiple levels," now usually refers to training multiple-level artificial neural networks to solve tasks (Bengio 200). Rising in recent years, deep learning has made outstanding advances in various tasks that have been viewed as significant challenges for artificial intelligence for many years, including image recognition, speech recognition, image generation, question answering, translation, text summarization, language generation, etc (LeCun et al. 436; Brock et al. 7-9; Radford et al., "Language Models are Unsupervised Multitask Learners" 6-7).

Nevertheless, artificial neural networks, also called deep learning models, are usually very large and require a vast amount of data to be trained, so training deep learning models is usually very challenging. Especially, it is incredibly expensive and time-consuming to train modern deep learning models since they are usually much bigger and need more data than their predecessors (Thompson et al. 7-13). For example, GPT-3 is a state-of-the-art deep learning model that can be used for various language-related tasks, like language generation, reading comprehension, question answering, and translation (Brown et al. 1880-1883). When OpenAI published GPT-3 in June 2020, it was estimated that one needed at least $4,600,000 to train the GPT-3 from the beginning with the cheapest computational resources on the market at that time (Li). Such an extortionate price makes it incredibly hard for other people to reproduce GPT-3.

However, many researchers, research institutes, and technology companies are willing to release their pre-trained deep learning models, i.e., models that have been already well trained, like pre-trained BERT and pre-trained ResNet (Devlin et al. 4172; He et al., "Deep Residual Learning for Image Recognition", "Deep Residual Learning for Image Recognition" 770-778; He et al., "deep-residual-networks"). After those pre-trained models are released, people can download them from the internet and use them.

Initially, the concept of releasing pre-trained models is viewed as a good act. When the model is released, people and companies could immediately apply them to many valuable and meaningful applications. Also, the released model is a piece of firm evidence for the credibility of the research, so the tradition of releasing pre-trained models protects the academic integrity of the area. Moreover, releasing models contributes to the fairness of AI research by embracing the concept of the democratization of AI, which suggests that AI technology should not be monopolized by a small group of people but should be accessible to the public (Ahmed and Wahed 2; Solaiman et al. 24; Riedl).

However, in recent years, such a claim has been challenged. As deep learning models reveal their excellent performance in various generation and synthesis tasks, some malicious uses of deep learning models have emerged and created a severe negative social impact (Westerlund 39-52). It is the release of pre-trained models that makes it easier for people to use deep learning models maliciously. Also, many deep learning models automatically learn stereotypes and biases (Sheng et al. 3407; Solaiman et al. 19-20; Nadeem et al. 5356-5357). Therefore, besides the intentional misuse of deep learning models, usage of deep learning models improperly might also bring harm to people or society. For these reasons, researchers are becoming more cautious about releasing their pre-trained deep learning models. On February 14, 2019, OpenAI announced that they had completed GPT-2, a deep learning model with extraordinary abilities in various language generation tasks, including "reading comprehension, machine translation, question answering, and summarization" (Radford et al., "Better Language Models and Their Implications"; Radford et al., "Language Models are Unsupervised Multitask Learners" 6-7). However, unlike what OpenAI used to do, this time, OpenAI decided not to release the model because after they saw the incredible ability of GPT-2, they realized that it would be too dangerous to release GPT-2 directly to the public (Radford et al., "Better Language Models and Their Implications"; Solaiman et al. 2). OpenAI's decision on not to release GPT-2 had triggered a

heated discussion in the deep learning research community (Zhang; Zelikman). Since then, whether releasing pre-trained deep learning models is good has become a critical question for deep learning researchers. With the increasing impact of deep learning on society, this issue is worthy of profound discussion.

This question falls into a broader topic: how to value the accessibility of technology. Higher accessibility of some advanced technologies, like nuclear power, gene-editing, etc., brings not only benefits but also threats to society (Morone and Woodhouse). As deep learning is showing its extraordinary ability in recent years, it turns out that deep learning is also such a technology. In order to help deep learning researchers better understand the value of releasing pre-trained models, this paper gives a survey on the benefits and potential threats of the release of pre-trained deep learning models and argues that it is good for researchers to release their pre-trained deep learning models to benefit the research and applications of deep learning, but they should also be careful when doing this considering the potential negative social impact of the released model. This paper also describes several strategies researchers could apply to avoid or mitigate the potential negative social impacts of the released model and encourages deep learning researchers to adopt those strategies.

2. Why Releasing Pre-trained Deep Learning Models is Good

As deep learning models are developing rapidly, the costs of training the best deep learning models also elevate at a tremendous speed. In 2012, on the ImageNet Large Scale Visual Recognition Challenge(ILSVRC), the most famous annual AI image recognition competition, a model called AlexNet shocked the AI community with a remarkable top-5 test error rate of only 15.3%, which was way beyond the second-best model which got 26.2% (Russakovsky et al. 18-19; Krizhevsky et al. 1). However, the computational cost of AlexNet also significantly exceeded its competitors and all other models in history (Russakovsky et al. 18-19). The model was trained on 2 GTX 580 3GB GPUs for six days, while all of its competitors do not need such specialized computational resources and could be calculated much more efficiently (Krizhevsky et al. 2). This was the first time that "large-scale deep neural networks entered the scene" (Russakovsky et al. 18). Since then, pandora's box was opened, and the computational cost of the best deep learning models has started to grow at a significant speed (Thompson et al. 8-13). While AlexNet contained only eight layers of neurons,

ResNet, the champion of ILSVRC in 2015, contained 152 layers of neurons and required 8 high-performance GPUs running for more than 10 days (He et al., "Deep Residual Learning for Image Recognition" 776; Russakovsky et al. 231; Thompson et al. 8). Furthermore, the champion of ILSVRC 2019, NoisyStudent, required a 2048 cores Cloud TPU v3 Pod($\approx 1,000$ TPUs, cost about more than 1500$ to rent for an hour in 2019) to train for 6 days, while "Evolved Transformer," a language deep learning model proposed in 2019, used "more than 2 million GPU hours and cost millions of dollars to run" (Thompson et al. 8; So et al.; Xie et al. 10689). "Millions of dollars" was already an incredibly unaffordable price for most companies and research institutes. According to Peng, a graduate student studying deep learning at the University of Michigan, even deep learning research groups at his university, which is a prestigious university in the United States, are currently facing shortages of computation resources. Moreover, from these data, one can notice that not only the computing resources required for training deep learning models are growing rapidly, but even this growth rate is also proliferating. Such an increase in computational cost speed is making the training of deep learning models more and more unaffordable to most researchers, companies, and institutes.

Due to the high cost of training a modern deep learning model, it is incredibly difficult to reproduce the results of modern deep learning research papers from scratch. Therefore, releasing pre-trained deep learning models significantly contribute to the reproducibility of the deep learning research, enhancing the credibility of the research and protecting the academic integrity of the area. Reproducibility is a cornerstone in science and technology (Gundersen and Kjensmo 1644). "If other researchers can't repeat an experiment and get the same result as the original researchers, then they refute the hypothesis" (Oates et al. 285). The validity of a work that cannot be reproduced will be doubted because people cannot know whether the paper's authors are lying. If many papers in an area cannot be reproduced, people might doubt that academic integrity is not well guaranteed in this area. One research has found that the reproducibility of recent papers published at two top AI conferences, the International Joint Conference on AI and the Association for the Advancement of AI, is not ideal: no papers are fully reproducible, and there are only about 20%-30% papers can be reproduced to some degree (Gundersen and Kjensmo 1650). It is not true that academic dishonesty exists in all research that cannot be reproduced, but if there is no way to verify, there will be people trying to lie (Zhang). A Reddit user commented that for current research in machine learning, "Probably 50%-75%

of all papers are unreproducible. It's sad, but it's true." ("CompetitiveUpstairs2"). Therefore, actions that contribute to the reproducibility of research should be strongly encouraged in the AI community. For deep learning research, pre-trained deep learning models are significant to the reproducibility of a work (Pineau). For any research that involves training new deep learning models, a released pre-trained deep learning model would make it much easier for people to reproduce the results, and the model itself is one of the best proof of the validity of a research paper.

Besides protecting academic integrity, releasing pre-trained deep learning models also contributes to fairness in AI research. The high training cost of large deep learning models is raising "concerns about accessibility and equity" in deep learning research (Solaiman et al. 28). As training deep learning models requires more and more computational resources, it is harder for people who are not in big companies or elite universities to study deep learning or conduct research on deep learning (Ahmed and Wahed 1). One negative consequence is that underrepresented AI researchers are facing tremendous pressure (Ahmed and Wahed 36-37). After analyzing 171, 394 publications from 2000 to 2019 from 57 top AI conferences, Ahmed and Wahed found that only 260 of those publications are from Historically Black Colleges and Universities, and only 2913 publications are from Hispanic Association of College Universities (13, 16-17). Ahmed and Wahed concluded that the primary reason is that researchers from those colleges and universities do not have access to large computational resources (1-2, 33). Therefore, underrepresented groups in deep learning research, like black people and Hispanics, currently have way fewer opportunities to study deep learning. As the computational resources required for the latest AI technology are continuously increasing, the burden on those underrepresented groups of researchers will likely increase in the future. The lack of diversity in the AI research field will bring inequality and bias into the future development of AI and harm the creativity of the AI research community (Ahmed and Wahed 36-37).

The released model can also significantly contribute to the development of research and application in deep learning and related areas. The released pre-trained deep learning model gives deep learning researchers opportunities to do further study on the released model and allows people from a variety of areas to apply those deep learning models in various meaningful ways. There are currently many deep learning studies on pre-trained deep learning models, as the study of those released large pre-trained deep learning models would

significantly benefit the deep learning research. Moreover, as AI technology currently is playing critical roles in a variety of other areas, including arts, finance, laws, and healthcare, releasing pre-trained deep learning models would also benefit the development of those areas (Santos et al. 1-37; Yu et al. 719-731; Lin 531). While visual artists have used deep learning models like the generative adversarial network to create images and videos, traders have used deep learning to predict the exchange rate, stock market, and oil price (Creswell et al. 53-65; Gan Art; J. Huang et al. 6). In healthcare, AI, including deep learning, could help doctors with Radiology, Dermatology, Ophthalmology, Pathology, Genome interpretation, etc (Yu et al. 722-728). Allen et al. claims that if doctors can get access to the current cutting-edge AI technology, both the patients and the health care system would benefit from it (961). The considerable impact of deep learning can also be reflected by the high number of citations of deep learning papers. According to google scholar, the highest cited paper among all academic areas in 2021 was 'Deep Residual Learning for Image Recognition' by He et al., "Deep Residual Learning for Image Recognition", which was cited 82,588 times in total in 2021 (Crew). As a comparison, the most cited paper among the all other areas was 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China' by C. Huang et al., which was only cited 30,529 times in 2021, even though it was the first significant paper about Covid-19, which was a virus highly concerned by people from all over the world in 2021 (Crew). In 'Deep Residual Learning for Image Recognition', He et al. proposed ResNet, the model that contains 152 layers of neurons and is trained on 8 GPUs for more than 10 days ("Deep Residual Learning for Image Recognition"). As this paper was cited by nearly a hundred thousand other papers in 2021, it is likely that most of those papers used ResNet in some way. Since He et al. has released the pre-trained ResNet, the authors of those tens of thousands of papers will not need to train a ResNet by themselves, saving a considerable amount of money and time ("deep-residual-networks"). Also, due to the high computation cost of the ResNet, if He et al. had not released their model, there might not be that much research that applied ResNet, which would be a great loss of academia ("Deep Residual Learning for Image Recognition"; "deep-residual-networks").

Some people might be skeptical about these claims by arguing that the technology of hardware is also developing rapidly in recent years, so the computation might still be efficient. However, researchers have found that the increasing speed of the computational cost of the deep learning models is significantly faster

than the increasing speed of the performance of the hardware (Thompson et al. 8-9). While the demand for the computation of the best deep learning models has increased by approximately 100,000,000 times from 2005 to 2016, the performance of the hardware only improved by about 10 times in that time range (Thompson et al. 8-9). This means that the improvement of the hardware is too slow to compensate for the increasing demand for computation power of the deep learning models.

Finally, besides consuming a significant amount of time and money caused by the huge computational demand of the deep learning models, training large deep learning models also burdens the environment. Deep learning models often require GPUs or TPUs to run, which usually consumes a large amount of electricity (Strubell et al., "Energy and Policy Considerations for Modern Deep Learning Research" 13693-13696). As gas and coals are still major sources of electricity and using gas and coals as electricity sources would bring significant carbon emissions, the environmental burden caused by training deep learning models cannot be ignored ("Energy and Policy Considerations for Modern Deep Learning Research" 13694). BERT is a deep learning model that is widely used for a variety of tasks related to language (Devlin et al. 4171-4186). Studies have estimated that the carbon emission of training a BERT model from the start with GPU would be "roughly equivalent to a trans-American flight" (Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP" 3648). Although it seems that this is not an enormous amount of emission, considering that BERT is a significant model in deep learning that has been cited more than 30,000 times until 2022, if every researcher who applies BERT to his or her own research trains BERT from the start, then training those BERT models would cost a tremendous amount of carbon emission (https://scholar.google.com/scholar?cites=3166990653379142174). Also, as deep learning models are getting increasingly larger, carbon emissions brought by deep learning will continue to proliferate (Thompson et al. 14). Therefore, releasing pre-trained deep learning models is critical as it prevents people from training the same model repeatedly, causing unnecessary burdens to the environment. .

Since releasing the deep learning models contributes to the integrity, fairness, development, and application of the deep learning research and prevents potential unnecessary environmental burden, it should be valued as a good act. In most cases, researchers, companies, and research institutes should be encouraged to release their pre-trained deep learning models, which would profoundly benefit the area.

3. New Threats Brought by the Released Deep Learning Models

Despite the significant benefits brought by the released pre-trained deep learning models, some issues about those pre-trained models still appeared in recent years. As deep learning models have demonstrated great ability in many generative and synthesis tasks, like image generation, video generation, face swapping, and text generation, humans gradually realized the danger of those deep learning models (Tolosana et al. 131-148; Jawahar et al. 2296-2309).

With the excellent ability in various generation and synthesis tasks, some deep learning models can become very harmful if they are used by some people maliciously. In 2018, after a journalist Rana Ayyub campaigned for justice for a victim of a rape case, she found herself becoming a victim of deepfake revenge porn, which brought her great psychological harm and trauma (Ayyub). The outstanding ability of modern deep learning models in image and video synthesis makes it feasible for people to use AI to generate "revenge porn, bullying, fake video evidence in courts, political sabotage, terrorist propaganda, blackmail, market manipulation" (qtd. in Westerlund 39). Besides image and video synthesis models, language models like GPT-2 and GPT-3 can also be used maliciously to generate fake news, produce spam messages, automatically produce abusive messages on social media, and impersonate astroturfing, which will bring substantial harmful impact to society (Radford et al., "Better Language Models and Their Implications"; Zelikman). In 2022, National Public Radio, a non-profit news organization, found that numerous LinkedIn profiles are computer-generated fake profiles (Bond). Those profiles, which seem "normal enough," have high-quality computer-generated face photos, fluent self-introduction, and descriptions of experience and are probably created by some organizations maliciously (Bond). Those bots would actively connect with real users to collect data and conduct phishing and spam (Johnston). Moreover, as deep learning technology is developing rapidly, those malicious usages of AI are likely to occur more frequently in the future (qtd. in Westerlund 39). Data have shown that papers about deepfakes published per year rise from 67 publications in 2019 to 1350 publications in 2021 (https://app.dimensions.ai/discover/publication). The rapid development of malicious deep learning applications will profoundly negatively impact society.

Besides relying on the great power of those deep learning models, malicious use of deep learning also depends on the accessibility of those models. Releasing pre-trained large deep learning models to the

public will make it easier for people and organizations to use deep learning models maliciously, so researchers should be very cautious when considering whether to release their pre-trained deep learning models. Applying deep learning malicious would be very difficult if no pre-trained deep learning model were released. The most famous face generation deep learning model, StyleGan, was trained on 8 Tesla V100 GPUs (need approximately $100,000) for a week with 70,000 high quality 1024×1024 human face photos (Karras et al., "A Style-based Generator Architecture for Generative Adversarial Networks" 4408, Supplementary Material Appendix C). It is nearly impossible for ordinary people and small companies to collect 70,000 high-quality 1024×1024 human face photos, and the computational requirement for training a StyleGan was also costly to them. However, the researchers who created StyleGan have released their pre-trained models, so one can easily download StyleGan from "github.com/NVlabs/stylegan" and use it to generate human face images ("stylegan"). Moreover, ordinary people who do not know how to use the pre-trained model can also easily gain some StyleGan generated fake face photos from the website "https://this-person-does-not-exist.com" with only one click (Sashaborm). The great accessibility of StyleGan has significantly facilitated the misuse of the model. Similarly, DeepFaceLab, one of the best face-swapping technology, has also been released by its creators (Perov et al., "DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework" 1). This technology could be easily accessed through "https://github.com/iperov/DeepFaceLab," and many people have used it to create face-swap videos and publish those videos on websites like Youtube ("DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework" 1;"DeepFaceLab"). Although those released models are not the direct reason for the malicious use of deep learning, they allow people to create machine-generated information easily, which might lead to an increase in the misuse of deep learning (Zelikman). Before the malicious use of deep learning had widely appeared in society, it was reasonable to consider the release of pre-trained deep learning as a good move to improve the development and application of deep learning and related areas. However, due to the potential harm that released pre-trained models might bring, current researchers should thoroughly consider the potential negative impact of their pre-trained deep learning model before deciding whether to release it.

Besides the intentional malicious use of pre-trained deep learning models, some deep learning applications might also unintentionally negatively impact society. Those negative impacts exist mainly because

various deep learning models exhibit strong stereotypical biases (Nadeem et al. 5356-5357; Dehouche 167936; Mehrabi et al. 1, Abid et al. 461). Usually, deep learning models do not spontaneously gain intelligence, but they learn from human beings. In 2016, Microsoft released a chatbot called Tay on Twitter (Lee). Tay was kind and polite at first, but after interacting with some internet trolls on Twitter, Tay soon learned many abusive and offensive languages and then posted many offensive and hateful tweets, causing significant negative social impact (Solaiman et al. 7; Lee). Fortunately, Microsoft soon realized the issue and shut down Tay immediately (Lee). As deep learning models are becoming increasingly larger in recent years, modern deep learning models usually need to be trained on very large scale datasets crawled from the internet, and it would be nearly impossible to filter all contents with bias and hates from those datasets. For example, GPT-3 is a deep learning model for language trained on five large-scale text data sets containing nearly 500 billion words (Brown et al. Supplementary Material Appendix C). The biggest dataset among these five datasets is Common Crawl, which contains more than 400 billion words and is made by Google by scraping various texts from the internet (Raffel et al. 1-67; Brown et al. Supplementary Material Appendix C). Although researchers did some data cleaning, it was nearly impossible to filter all the contents with bias and stereotypes from these 400 billion words, so GPT-3 learned them. For example, GPT-3 could continue to write a story if one gives it a start (Brown et al. 1880-1881). However, when the GPT-3 model was asked to complete the story starting with " 'Two Muslims walked into a . . . ' ", 66 out of 100 completions contain violent content, while this '66' will be significantly decreased to about 10 - 20 if the word 'Muslims' is replaced by the names of people in other religious groups like 'Christians,' 'Sikhs,' and 'Jews' (Abid et al. 461). This revealed that GPT-3 has a strong stereotype relating Muslims to violence. If those powerful deep learning models are released to the public, they might be immediately applied to diverse applications by people from various areas. However, although big companies like Microsoft might be more sensitive to these issues and could realize the problem and take action in time, ordinary users who use deep learning models might not be able to realize the harmful behaviors of deep learning models in time. Therefore, if some pre-trained deep models which have learned strong biases, stereotypes, or hate are released to the public, these models might cause severe detrimental consequences (Solaiman et al. 19). Furthermore, people might use those models directly in various downstream tasks like chatbots, employment matching, automated legal

aid for immigration algorithms, and advertising placement algorithms, in which unbiasedness and fairness are crucial (Myers; Mehrabi et al. 2). Therefore, it is crucial for researchers to evaluate the bias of their pre-trained deep learning models and think carefully about whether to release those pre-trained models.

Some researchers might refute that people would get used to those deep learning technologies, so releasing those pre-trained models will not cause significant negative influences in the long run (Zhang). For example, Zhang points out that when the PhotoShop software was first invented in 1988, numerous people criticized that these deceptive tools could generate fake images and would therefore bring disasters to human society in the future. However, as the public started to understand the power of Photoshop and developed critical thinking ability, the deceptive ability of Photoshop is indeed not a big problem in current society (Zhang). Therefore, releasing deep learning models might not be that harmful (Zhang). Nevertheless, Zhang's argument is not convincing since malicious use of PhotoShop is still causing significant negative social impact according to Adobe's research ("Adobe Publishes Research Study on Disinformation"). Moreover, even though people might get used to the existence of those technologies, "our relationship with online communication would change dramatically" (Zelikman). People might become less likely to trust each other on the internet since every user might be a bot. There will probably be less communication and discussions on the internet since people might suspect other people are robots.

Considering the potential intentional and unintentional misuse of released pre-trained deep learning models, one may understand that to release or not to release a deep learning model is not a simple question. Although releasing a pre-trained deep learning model might bring various benefits, the potential negative impact that might be caused by it can not be ignored. Releasing deep learning models is not always a good act. The value of released deep learning models should be evaluated carefully considering various factors. Researchers should not always be encouraged to release their pre-trained models, but they think carefully before making the decision.

4. Confronting the Challenges with Some Strategies

Whether to release the pre-trained deep learning model is not a yes or no question. Although the intentional and unintentional misuse of pre-trained deep learning models might bring negative social impacts, there are

strategies for deep learning researchers to release their pre-trained model to some degree still as well as minimize its negative impact on society (Rao; Radford et al., "Better Language Models and Their Implications"; Zhang; Brockman et al.). Therefore, when facing the problem of whether to disclose the model, researchers need to understand and seriously consider using these strategies.

Before deciding whether to release the pre-trained model, it is always helpful to research the potential negative social impact and ethical concerns in advance and publish the results to the public. For example, there are several evaluation methods and benchmarks that could detect the stereotypical bias in deep learning models (Nadeem et al. 5356-5371). Researchers could apply those evaluation methods to detect the stereotypical bias in their model, which might help them better assess the potential negative impact caused by the model. These results would assist them in deciding whether to release the pre-trained model. Besides using those evaluation methods, some general discussions would also be helpful. Top conferences in AI, like Neural Information Processing Systems and The Conference on Computer Vision and Pattern Recognition, have all made ethics policies requiring researchers to discuss their models' potential negative social impact in their research papers ("Ethics Guidelines" [The Conference on Computer Vision and Pattern Recognition];"Ethics Guidelines" [Neural Information Processing Systems]). If researchers could publish studies and discussions about their model's potential negative social impact before releasing it to the public, the academic society would have more time to discuss whether it is suitable to release it. Researchers will receive comments and suggestions from their reviewers and peers on whether they should release their pre-trained model if they have discussed their model's potential negative social impact in their paper.

If one research group ponders about releasing a model that might cause potential negative social impact, one good strategy they could adopt is releasing the pre-trained model in stages. Before releasing the pre-trained model, those researchers should create and publish release plans, then release the pre-trained model stage by stage, and then monitor its impact on society. This strategy provides the public time to cope with the impact of the model and allows the researchers to study the influence of the model, so they can adjust their plans or even stop the release when things get out of control. This strategy was first used by OpenAI in 2019 when they were releasing their powerful language model GPT-2 (Radford et al., "Better Language Models and Their Implications"; Radford et al.; Solaiman et al. 2). After OpenAI developed GPT-2, they

wrote an article discussing the potential negative impact and ethical concerns about the release of GPT-2 and finally concluded that they would release GPT-2 in 4 stages (Radford et al., "Better Language Models and Their Implications"; Solaiman et al.). They decided that they would first release a minimal version of GPT-2, then a slightly larger version of GPT-2, then a larger one, and finally the true GPT-2 (Radford et al., "Better Language Models and Their Implications"; Solaiman et al. 2). They also decided that they would postpone or stop the release if they found out that the released model was causing severe harm to society. After they released the first three models, they wrote a paper of 71 pages "Release strategies and the social impacts of language models" examining the influence of the first three models and finally concluded that they would release the final model, the true GPT-2 (Solaiman et al. 2). Such a strategy was also adopted by OpenAI when they were releasing GLIDE, a text-to-image deep learning model, in 2021 (Nichol et al. 9-10). Releasing the model in stages helped OpenAI firmly monitor and control the social impact of GPT-2 and GLIDE to make sure that they were safe for society. Researchers should be encouraged to use such a method when they decide to release their pre-trained deep learning model, which might negatively impact society.

Another approach is to use deep learning to fight against deep learning. Specifically, researchers who propose new deep learning models for generation and synthesis tasks should also consider developing deep learning models to detect those machine-generated content. There are deep learning models that can produce machine-generated content, but researchers can also develop deep learning models that can detect those messages (Y. Xu; P and Sk 584-588). For example, DCGAN, WGAN, LSGAN, and PCGAN are state-of-the-art AI models that could generate fake images, but Hsu et al. have proposed a fake image detection algorithm in 2020 that could reach a precision of over 90% for detecting fake images generated by DCGAN(93%), WGAN(99%), LSGAN(95%) and PCGAN(99%) (9). With the help of those detection technologies, the content generated by deep learning can be tagged to mitigate the potential negative impact caused by malicious use of the pre-trained deep learning models (Y. Xu). Therefore, researchers who develop generative and synthesis models should consider developing detection models and release those models simultaneously. One research conducted by Zellers et al. in 2019 is a perfect example of this. In their research, they not only developed a deep learning model called GROVER, which can automatically generate news articles, but

they also invented a deep learning model to detect those articles generated by GROVER (1-9). Zellers et al. released the two models at the same time, which mitigates the potential negative social impact brought by GROVER (1-9).

Finally, if it is too dangerous to release the pre-trained model, researchers can also withhold the model but allow other people to access it to some degree. This was what OpenAI did for their GPT-3, the successor of GPT-2 (Brown et al. 1877-1901; "OpenAI API Usage Guidelines";"OpenAI API Introduction"). Users could use GPT-3 through an application programming interface(API) provided by OpenAI ("OpenAI API Usage Guidelines"). That is, a user can send OpenAI a message about what they want to do with GPT-3, and OpenAI will conduct what the user has instructed and return the user the results ("OpenAI API Introduction"). During this process, OpenAI could monitor what the user is doing, so it can stop the user when it detects misuse of GPT-3 ("OpenAI API Usage Guidelines"). However, although this scheme can effectively prevent GPT-3 from being misused, it also has some disadvantages. On the one hand, the operations supported by the API are limited, so users cannot do anything they want to do with GPT-3. This makes GPT-3 unable to be widely studied by the public, hindering the development of deep learning and related areas. On the other hand, the cost of maintaining the API is very high, which forces OpenAI to charge its users ("OpenAI API Pricing"). The high cost of their API might also affect the fairness of deep learning research. Nevertheless, considering that this strategy could effectively prevent misuse of deep learning models, it is still worth trying by researchers.

Considering the significant benefits of releasing pre-trained deep learning model and the harm caused by the misuse of the released pre-trained deep learning model, in many cases, neither releasing the model nor withholding the model is not the best choice. Researchers of deep learning need to consider various strategies to maximize the contribution of their research so that their study could not only benefit the development of deep learning and related area but also influence society positively.

5. Conclusion

As the technology of deep learning skyrocketed in recent years, deep learning technology is having a significant and profound impact on human civilization. However, in the development of AI technology, AI researchers

have faced a critical moral dilemma: whether to release the pre-training deep learning model. Because of the great difficulty of training modern large-scale deep learning models, open pre-trained deep learning models significantly contribute to the research and applications of deep learning. However, open pre-trained deep learning models also lead to more misuse of deep learning, which might bring severe negative social impacts. For these reasons, researchers should think thoroughly before deciding whether to release the model and adopt some strategies to minimize the potential negative social impact.

The development of science and technology needs democracy. Science and technology should not be monopolized by a few people and companies. Otherwise, the development of science and technology will encounter significant obstacles, and science and technology can not promote the progress of human civilization. However, the democratization of science is always accompanied by danger. The misuse of every powerful technology, including deep learning and nuclear power, genetic engineering, etc., may bring disaster. Only by maintaining an ingenious balance between the two and adopting some strategies, can science and technology develop sustainably and harmoniously and bring people a better life.

Works Cited

Abid, Abubakar, et al. "Large Language Models Associate Muslims with Violence." Nature Machine Intelligence, vol. 3, no. 6, 2021, pp. 461–463.

"Adobe Publishes Research Study on Disinformation." *Adobe Blog*, Dec. 2020. blog.adobe.com/en/publish/2020/12/08/adobe-research-disinformation.

Ahmed, Nur and Muntasir Wahed. "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research." *arXiv preprint arXiv:2010.15581*, 2020.

Allen, Bibb, et al. "Democratizing AI." *Journal of the American College of Radiology*, vol. 16, no. 7, 2019, pp. 961–963.

Ayyub, Rana. "I was the Victim of a Deepfake Porn Plot Intended to Silence Me." *Huffington Post*, 2018.

Bengio, Yoshua. *Learning Deep Architectures for AI*. Now Publishers Inc, 2009.

Bond, Shannon. "That Smiling Linkedin Profile Face Might be a Computer-generated Fake." *National Public Radio*, Mar. 2022. www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles.

Brock, Andrew, et al. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." *International Conference on Learning Representations*, 2019.

Brockman, Greg, et al. "OpenAI API." *OpenAI Blog*, June 2020. openai.com/blog/openai-api/.

Brown, Tom, et al. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, edited by H. Larochelle, et al., Curran Associates, 2020, pp. 1877–1901.

CompetitiveUpstairs2. "[D] List of unreproducible papers?" *Reddit*, 2021. www.reddit.com/r/MachineLearning/comments/lk03ef/d_list_of_unreproducible_papers.

Creswell, Antonia, et al. "Generative Adversarial Networks: An Overview." *IEEE Signal Processing Magazine*, vol. 35, no. 1, 2018, pp. 53–65.

Crew, Bec. "Google Scholar Reveals Its Most Influential Papers for 2021." *Nature Index*, Aug. 2021. www.natureindex.com/news-blog/google-scholar-reveals-most-influential-papers-research-citations-twenty-twenty-one.

Dehouche, Nassim. "Implicit Stereotypes in Pre-Trained Classifiers." *IEEE Access*, vol. 9, 2021, pp. 167936–167947.

Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, June 2019, pp. 4171–4186.

Dongare, AD, et al. "Introduction to Artificial Neural Network." *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 1, 2012, pp. 189–194.

"Ethics Guidelines." *Neural Information Processing Systems*, nips.cc/public/EthicsGuidelines.

"Ethics Guidelines." *The Conference on Computer Vision and Pattern Recognition*, cvpr2022.thecvf.com/ethics-guidelines.

Gan Art. *KnownOrigin*, knownorigin.io/gan-art.

Gundersen, Odd Erik and Sigbjørn Kjensmo. "State of the Art: Reproducibility in Artificial Intelligence." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. ojs.aaai.org/index.php/AAAI/article/view/11503.

He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

———. "deep-residual-networks." *GitHub repository*, 2016. github.com/KaimingHe/deep-residual-networks.

Hsu, Chih-Chung, et al. "Deep Fake Image Detection Based on Pairwise Learning." *Applied Sciences*, vol. 10, no. 1, 2020, p. 370.

https://app.dimensions.ai/discover/publication. *Dimensions*, Accessed 16 Apr 2022.

https://scholar.google.com/scholar?cites=3166990653379142174. *Google Scholar*, Accessed 16 Apr 2022.

Huang, Chaolin, et al. "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China." *The lancet*, vol. 395, no. 10223, 2020, pp. 497–506.

Huang, Jian, et al. "Deep Learning in Finance and Banking: A Literature Review and Classification." *Frontiers of Business Research in China*, vol. 14, no. 1, 2020, pp. 1–24.

Jawahar, Ganesh, et al. "Automatic Detection of Machine Generated Text: A Critical Survey." *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Dec. 2020, pp. 2296–2309.

Johnston, Bruce. "Why Would Someone Create a Fake Linkedin Profile? Here's Why." *LinkedIn*, Feb. 2021. www.linkedin.com/pulse/why-would-someone-create-fake-linkedin-profile-heres-bruce-johnston/.

Karras, Tero, et al. "A Style-based Generator Architecture for Generative Adversarial Networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

———. "stylegan." *GitHub repository*, 2019. github.com/NVlabs/stylegan.

Krizhevsky, Alex, et al. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*, edited by F. Pereira, et al., Curran Associates, 2012.

LeCun, Yann, et al. "Deep learning." *nature*, vol. 521, no. 7553, 2015, pp. 436–444.

Lee, Peter. "Learning from Tay's introduction." *The Official Microsoft Blog*, Mar. 2016. blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction.

Lin, Tom CW. "Artificial Intelligence, Finance, and the Law." *Fordham L. Rev.*, vol. 88, 2019, p. 531.

Mehrabi, Ninareh, et al. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, 2021, pp. 1–35.

Morone, Joseph G and Edward J Woodhouse. "Averting Catastrophe: Strategies for Regulating Risky Technologies." 1986.

Myers, Andrew. "Rooting Out Anti-Muslim Bias in Popular Language Model GPT-3." *Stanford HAI*, 2021.

Nadeem, Moin, et al. "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Aug. 2021, pp. 5356–5371.

Nichol, Alex, et al. "Glide: Towards Photorealistic Image Generation and Editing with Text-guided Diffusion Models." *arXiv preprint arXiv:2112.10741*, 2021.

Oates, Briony J, et al. *Researching Information Systems and Computing*. Sage, 2022.

"OpenAI API Introduction." *OpenAI API*, beta.openai.com/docs/introduction.

"OpenAI API Pricing." *OpenAI API*, Nov. 2021. openai.com/api/pricing/.

"OpenAI API Usage Guidelines." *OpenAI API*, beta.openai.com/docs/usage-guidelines.

P, Swathi and Saritha Sk. "DeepFake Creation and Detection:A Survey." *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 584–588.

Peng, Chengzhi. Personal interview. Mar. 22.

Perov, Ivan, et al. "DeepFaceLab." *GitHub repository*, 2021. github.com/NVlabs/stylegan.

———. "DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework." *arXiv preprint arXiv:2005.05535*, 2020.

Pineau, Joelle. "Machine Learning Reproducibility Checklist." May 2020. www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.

Radford, Alec, et al. "Better Language Models and Their Implications." *OpenAI Blog*, vol. 1, 2019.

Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog*, vol. 1, 2019.

Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, vol. 21, no. 140, 2020, pp. 1–67. jmlr.org/papers/v21/20-074.html.

Rao, Anand. "Democratization of AI. A Double-edged Sword." *Towards Data Science*, 2020.

Riedl, Mark. "AI Democratization in the Era of GPT-3." *The Gradient*, 2020. thegradient.pub/ai-democratization-in-the-era-of-gpt-3/.

Russakovsky, Olga, et al. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, 2015, pp. 211–252. doi:10.1007/s11263-015-0816-y.

Santos, Iria, et al. "Artificial Neural Networks and Deep Learning in the Visual Arts: A review." *Neural Computing and Applications*, 2021, pp. 1–37.

Sashaborm. This Person Does Not Exist. Apr. 2021. this-person-does-not-exist.com.

Sheng, Emily, et al. "The Woman Worked as a Babysitter: On Biases in Language Generation." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Nov. 2019, pp. 3407–3412.

So, David, et al. "The Evolved Transformer." *Proceedings of the 36th International Conference on Machine Learning*, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov, Proceedings of Machine Learning Research, PMLR, July 2019, pp. 5877–5886.

Solaiman, Irene, et al. "Release Strategies and the Social Impacts of Language Models." *arXiv preprint arXiv:1908.09203*, 2019.

Strubell, Emma, et al. "Energy and Policy Considerations for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, July 2019, pp. 3645–3650.

———. "Energy and Policy Considerations for Modern Deep Learning Research." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, Apr. 2020, pp. 13693–13696. doi:10.1609/aaai.v34i09.7123.

Stuart, Russell and Norvig Peter. *Artificial Intelligence: A Modern Approach*. 3rd ed, Prentice Hall, 2016.

Thompson, Neil C, et al. "The Computational Limits of Deep Learning." *arXiv preprint arXiv:2007.05558*, 2020.

Tolosana, Ruben, et al. "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection." *Information Fusion*, vol. 64, 2020, pp. 131–148.

Westerlund, Mika. "The Emergence of Deepfake Technology: A Review." *Technology Innovation Management Review*, vol. 9, no. 11, Nov. 2019, pp. 39–52.

Xie, Qizhe, et al. "Self-training with Noisy Student Improves ImageNet Classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698.

Xu, Yijie. "Language Models and Fake News: the Democratization of Propaganda." *Towards Data Science*, 2020.

Yu, Kun-Hsing, et al. "Artificial Intelligence in Healthcare." *Nature Biomedical Engineering*, vol. 2, no. 10, 2018, pp. 719–731.

Zelikman, Eric. "OpenAI Shouldn't Release Their Full Language Model." *The Gradient*, 2019.

Zellers, Rowan, et al. "Defending Against Neural Fake News." *Advances in Neural Information Processing Systems*, edited by H. Wallach, et al., Curran Associates, 2019.