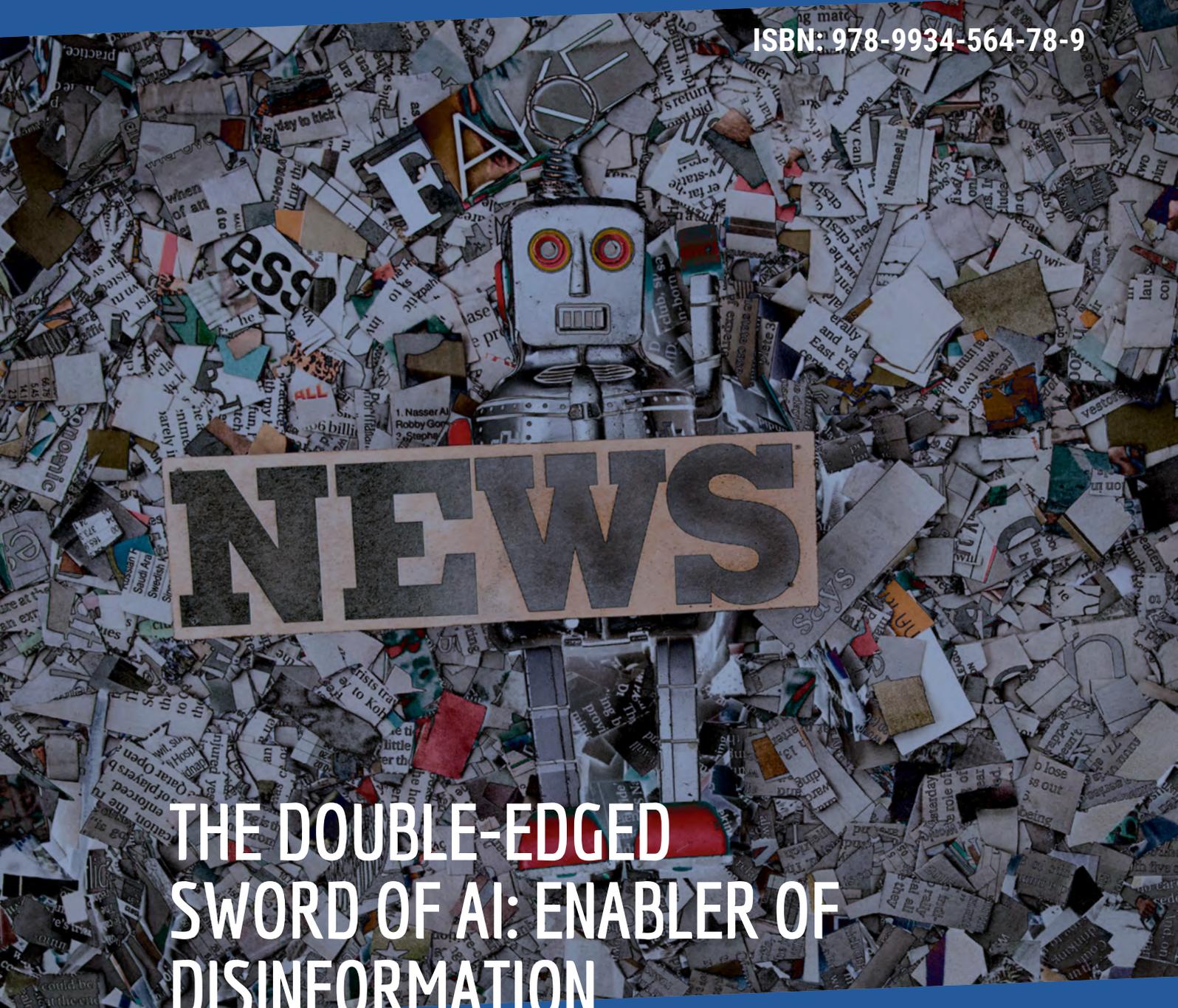


ISBN: 978-9934-564-78-9



NEWS

THE DOUBLE-EDGED
SWORD OF AI: ENABLER OF
DISINFORMATION

Published by the
NATO Strategic Communications
Centre of Excellence



ISBN: 978-9934-564-78-9

Authors: Alfonsas Juršėnas, Kasparas Karlauskas, Eimantas Ledinauskas, Gediminas Maskeliūnas, Julius Ruseckas

Contributors: Donatas Randomanskas, Luke Urbonavicius

Project Manager: Gundars Bergmanis-Korāts

Design: Kārlis Ulmanis

Riga, December 2021

NATO STRATCOM COE

11b Kalnciema iela

Riga LV1048, Latvia

www.stratcomcoe.org

[Facebook/stratcomcoe](https://www.facebook.com/stratcomcoe)

Twitter: [@stratcomcoe](https://twitter.com/stratcomcoe)

The authors would like to thank Mr Lukas Andriukaitis, Mr Viktoras Daukšas, Ms Alyssa Kann, and Mr Kamil Mikulski for sharing their expert insights into the field of disinformation. Of course, any misinterpretations and errors are entirely the fault of the authors.

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.

© All rights reserved by the NATO StratCom COE. Reports may not be copied, reproduced, distributed or publicly displayed without reference to the NATO StratCom COE. The views expressed here do not represent the views of NATO.

Abstract

The tendency to consume news on social media platforms has greatly increased over the last decade. Information can now be disseminated quickly, cheaply, and with easy access for consumers; this has rapidly boosted decentralized news production, often without editorial oversight. Adversarial agents are exploiting this situation to spread disinformation. Over the past ten years, the field of Artificial Intelligence (AI)/Machine Learning (ML) has experienced unprecedented growth in the development of applications for the automation of text, and the recognition and generation of visual and audio data. Do these burgeoning AI capabilities boost the abilities of malicious actors to manipulate crowds? AI now plays a vital role in generating synthetic content and enables the efficient micro-targeting used on social media platforms to spread disinformation messages, including hyper-realistic synthetic images, videos, audios, and text. This rather technical article has been written to inform practitioners, policymakers, and AI enthusiasts in NATO about how AI/ML technologies can be used to shape disinformation.



Table of contents

Introduction	6
Content preparation	9
Automated web data scraping	9
Automatic generation of text	10
Deepfakes: images, audio, video	12
Spreading disinformation	17
Bots, sock puppets, cyborgs	17
Disinformation enabled by social networks and data brokers	19
Possible future trends for AI-powered disinformation	21
Conclusions	24
Literature	27



” The current progress in AI [...] raise concerns that technology-enabled disinformation campaigns could amplify existing societal divisions, reduce trust in democratic institutions, and create other harmful outcomes.

Introduction

In the last decade, tremendous progress has been made in the field of Artificial Intelligence (AI), mainly due to the success of deep neural networks at performing various kinds of tasks. The idea of artificial neural networks originated in the 1940s.¹ However, artificial neural networks gained popularity, when developments in graphics processing units (GPUs) enabled the efficient training of otherwise computationally expensive, neural networks. In 2012 these advances made it possible for a network model called AlexNet to win the ImageNet Large Scale Visual Recognition Challenge.^{2,3} After this achievement in computer vision, the popularity of deep neural networks proliferated in other domains such as natural language processing, audio processing, and reinforcement learning.

Since this breakthrough, artificial neural networks have been consistently outperforming classic ML algorithms. In 2014, Facebook researchers published their work on the DeepFace model, which demonstrated substantial improvements in the accuracy of state-of-the-art facial recognition.⁴ In 2016, the AlphaGo program developed by Google subsidiary DeepMind Technologies was the first computer program to defeat a human champion at the strategy game Go.⁵ In 2018 and 2020, DeepMind used the protein structure

prediction systems AlphaFold 1 and AlphaFold 2 to win the Critical Assessment of protein Structure Prediction (CASP) competition, significantly advancing the state of the art.^{6,7} In 2019, OpenAI and DeepMind demonstrated AI programs that beat the best human players at the highly complex video games Dota 2⁸ and Starcraft 2.⁹ In 2020, OpenAI presented GPT-3, an autoregressive language model with a capacity of 175 billion parameters, which performed well on many different natural language processing tasks without any fine-



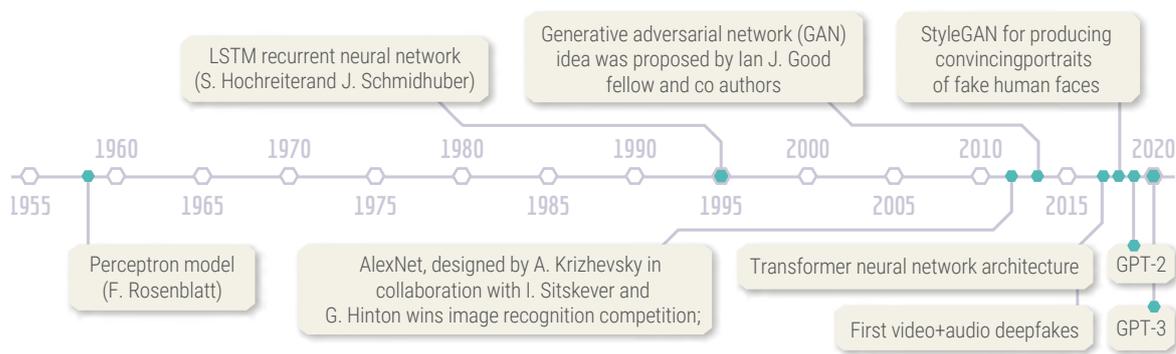


Figure 1. Some AI milestones relevant to the automated generation of disinformation content.

tuning.¹⁰ In 2021, OpenAI introduced DALL-E, a model that can generate images from textual descriptions in natural language.¹¹ Such rapid progress has enabled the development of huge, billion-parameter, general-purpose (sometimes referred to as task-agnostic) AI models to outperform their smaller predecessors.

These many examples demonstrate the increasingly significant role of AI in many different arenas. Unfortunately, disinformation is no exception. AI models can be used to generate malicious and compelling synthetic content (video, audio, text), simulate organic user comments and simple conversations, spread disinformation across social networks using bots, and exploit the micro-targeting capabilities of social platforms.

According to one recent study,¹² the general public is concerned with the trustworthiness of AI. The study highlights the importance of increasing transparency

about the underlying algorithms used to manage social media platforms, about data collection practices, and about control over the collected data; it also demonstrates the need for a better overall understanding of the security implications of data- and AI-driven services. Trustworthiness and transparency are crucial for automated systems that can flag news stories as “misleading” or “unreliable”, especially as third-party fact-checkers have a little impact on whether readers perceive the headlines they come across as true and accurate.¹³ The current lack of trustworthiness and transparency allows disinformation actors to spread manipulated or fake content efficiently and with little resistance.

The current progress in AI combined with the widespread use of social media raise concerns that technology-enabled disinformation campaigns¹⁴ could amplify existing societal divisions, reduce trust in democratic institutions, and create other harmful outcomes.^{15,16}



One of the main limitations of the methods currently in use is suboptimal data efficiency. Compared to humans, AI systems require huge volumes of data to learn even simple tasks. Developments in semi-supervised learning may influence the amount of data needed; pretraining general representation models on large amounts of unlabeled data and then using these models to learn specific tasks could potentially result in significantly faster rates of learning. The biggest shortcoming of current AI models seems to be a lack of common sense and reasoning capabilities. Despite recent advancements in the performance of various large-scale models, increasing model size and switching to more powerful hardware does not address these shortcomings. Fundamental algorithmic and methodological breakthroughs will be necessary to meet these challenges.

This article has adopted the broad definition of AI proposed by the European Commission in its *Artificial Intelligence Act*.¹⁷ An “Artificial Intelligence system” (AI system) is software developed using one or more of the following techniques or approaches:

- machine learning approaches, including supervised, unsupervised, and reinforcement learning that use a wide variety of methods, including deep neural networks
- logic- and knowledge-based approaches, including knowledge representation, inductive (logic)

programming, knowledge bases, inference, and deductive engines, (symbolic) reasoning and expert systems

- statistical approaches, Bayesian estimation, and search and optimisation methods

We aim to identify the potential risks posed by AI-enabled disinformation activity, and outline the challenges and limitations malicious actors must face. First, we discuss the preparation of disinformation content using web-scraped data and the state of the art for automatically generated content such as deepfakes, synthetic text, synthetic audio, etc. Then we explore the techniques and strategies used to spread disinformation. Finally, we present our conclusions and discuss likely future trends in the role of AI in disinformation campaigns from the perspective of the malicious actor.



Content preparation

Malicious actors interact with their target audiences through the user interface, just like genuine users of social media platforms do, by adding likes, posting comments, sharing opinions, and disseminating stories. To achieve their ends, malicious actors carefully prepare content for dissemination according to the disinformation narratives they wish to spread, and then use inauthentic accounts or curated bot networks to engage in authentic social networks as opinion leaders or opinion supporters. Artificial Intelligence can play a crucial role in creating such content^{18,19,20} through web scraping or more sophisticated algorithmic approaches that can be used to generate synthetic but realistic articles, comments, and other forms of engagement.

Automated web data scraping

Automatic content scraping is a multi-purpose activity and makes it possible to obtain large amounts of human-generated content in the form of text, images, video, and audio. Such data can help AI systems mimic organic behavior by, for example, using media bots to post fragments of organic content scraped from the web. More sophisticated malicious actors can use such data to train generative models to mimic scraped real-world data in various forms.

Furthermore, disinformation actors take advantage of machine translation services that can automatically translate the scraped content (e.g., content from websites that propagate radical ideologies or divisive narratives) into the language of a target audience. However, this method is currently reliable only for major languages where sufficiently accurate machine translation is available.²¹ In addition, numerous speech-

to-text services automatically create rough transcripts of audio recordings, making it possible to extract text from audio; malicious actors can use this method to spread polarizing speeches in text.

Despite the available models, services, and technologies, web content scraping is not a trivial task. Some websites have adopted the practice of dynamically changing their HTML code²² to cause errors in simple rule-based scrapers fine-tuned to extract specific parts of the HTML structure. This approach works because such random changes in underlying code can be calibrated not to alter the appearance of a website in the browser but cause scraping scripts to fail. Web-scraping enthusiasts must also deal with counter-scraping methods such as blocking user IPs after the detection of anomalous activity or a failed CAPTCHA screening.²³ Thanks to AI, however, such countermeasures may no longer be as effective. ML models can now create more robust scrapers with



the ability to compensate for alterations in HTML code.²⁴ Some commercial services²⁵ use machine learning tools to develop smart proxy rotation patterns, parse websites, and create simulated user fingerprints²⁶ to avoid detection by the automation detection algorithms of the websites they target. The complexity of CAPTCHA tests²⁷ has been gradually increasing due to such AI-based CAPTCHA targeting.²⁸

It seems evident that AI algorithms can improve existing web crawlers by making them robust to alterations of website code, and mimic human-like activity patterns to avoid automation detectors. High-quality web scraping enables various commercial services (for example, commercial brand monitoring, financial or sports statistics aggregation and many more) for the general public and makes it possible to prepare high-quality datasets for training ML models. However, web scraping is also a cheap way of obtaining large amounts of human-generated content that can be used to mask the activity of bot networks, spread divisive content, and prepare large datasets for abusive text generating models.

Automatic generation of text

Language modeling is done using various statistical techniques that determine the probability of a given sequence of words occurring in a sentence. After training in a certain language, a model can evaluate the most probable word or

symbol that might continue a given bit of text, usually called a “prompt”. When applied recursively on the original prompt and iterating over newly generated output, model can generate longer texts, often with questionable coherence. In recent years the capabilities of language modeling programs have significantly increased due to the development of attention-based transformer architectures such as GPT²⁹ and BERT.³⁰ In 2019, Radford et al. published their work on the Generative Pre-trained Transformer GPT-2,³¹ a language model that can autonomously generate coherent human-like paragraphs of text in response to input of just a single short sentence. The GROVER model for studying and detecting neural fake news came out the same year; it recognises synthetic text efficiently and can also efficiently and effectively generate multi-field documents such as journal articles. CTRL,³² a conditional language model, uses control codes to generate text in a specific style, with pre-determined content and task-specific behavior. Some alternative approaches using a variational autoencoder for text generation³³ exist in the literature, but the autoregressive architectures mentioned above are far more popular and better studied.

In 2020, OpenAI unveiled a sophisticated deep-learning-based language model called GPT-3,³⁴ an improved iteration of the GPT-2 language model mentioned above. The following section focuses on GPT-3 as other state-of-the-art methods share similar limitations.



GPT-3 implications and limitations

GPT-3 demonstrates state-of-the-art performance of various natural language processing (NLP) tasks (e.g., language modeling, question answering, summarization, machine translation, etc.) without parameter fine-tuning.³⁵ Users must provide only a few textual examples of the desired task to train the model. This brings us to the topic of a recently proposed paradigm called Few-Shot learning (FSL),³⁶ which corresponds to a family of pre-trained models able learn a specific task from only a few examples. Compared to earlier models and frameworks, GPT-3 does not require expert knowledge in AI. Built as task-agnostic, this tool can help malicious actors to create moderate- to high-quality messages at a much more impressive scale than ever before.³⁷

According to McGuffie and Newhouse,³⁸ GPT-3 shows a significant improvement over its predecessor, GPT-2, in generating biased text content, including hate speech in the context of disinformation. GPT-2 required elaborate fine-tuning on specific text corpora to generate realistic ideological propaganda; this is not the case for GPT-3. The model has been trained on nearly all available scraped web text, including content published by extremist communities such as QAnon, Atomwaffen Division, or the Wagner Group. This model was even able to reproduce the style and slogans of their disseminated content.

Despite its impressive performance, GPT-3 and other large language models do not have general intelligence or, shall we say, common sense. These models do not yet understand the meaning of words and still perform worse in reading comprehension and reasoning tasks compared to humans. Moreover, the text generated sometimes contradicts itself when multiple paragraphs are produced.³⁹ Interestingly, Brown et al.⁴⁰ show that, for some tasks, performance can improve sharply with a larger model, more parameters, and a bigger training dataset. This may lead to improvements in performing tasks at which GPT-3 currently fails. However, the opaque, black-box nature of language models creates another limitation: if a model generates text with unwanted biases or false ideas, it is difficult to locate and fix the problem.⁴¹

Large-scale models such as GPT-3 require significant amounts of data and computing resources, which gives an advantage to big tech companies. Retaining private ownership of the source code and the models they train also allows the tech companies to own state-of-the-art services⁴² based on these models. The general public and commercial partners can thus benefit from these cutting-edge AI models via the Application Programming Interface. However, since the publication of the GPT-3 article, multiple similar projects⁴³ have been undertaken (e.g., in China⁴⁴ and in Russia).⁴⁵ Unfortunately, models like GPT-3 are prone to generating texts using hateful, sexist, and racist phrases.^{46,47} Apart from integrated



service usage monitoring systems and the previously mentioned restrictions included in terms and conditions agreements, it is unclear how tech companies will monitor, regulate, and forbid the malicious use of few-shot task learning services after granting access to the general public.⁴⁸

Other NLP models and services

Hugging Face⁴⁹ provides many pre-trained transformer models that perform NLP tasks, such as text generation, in more than 100 languages. The model zoo includes BERT, GPT-2, and Google's sophisticated T5 model⁵⁰ and the multilingual version mT5. T5's architecture is designed as a unified framework that performs various language tasks (translation, question answering, classification, etc.) in a text-to-text format. The focus of another model, Nvidia Riva,⁵¹ is on conversational text generation and question answering. This tool would be more relevant for generating bot comments, not separate posts or articles.

Deepfakes: images, audio, video

If models such as GPT-3 can generate human-like written text, then recent developments in deep learning have produced generative models that can put those words in someone's mouth via deepfake technology. Early generative models amazed the general public with their ability to synthesize ultra-realistic imitations of human faces, today's models

create realistic, high-quality audio and video impersonations.

Fake images

Among the most frequent uses of AI encountered by social media analysts are AI-generated profile pictures.⁵² For example, in 2019 a bot swarm created hundreds of profile pictures for a network of fake accounts on Facebook, using images probably generated by StyleGAN2, which can be easily accessed on thispersondoesnotexist.com.

The best-known methods for generating fake images are Variational Autoencoders (VAEs)⁵³ and Generative Adversarial Networks (GANs).⁵⁴ These ML algorithms teach a neural network to produce new realistic images based on large amounts of real-world data that have been used to train them in various implementations.

VAEs first encode, then decode. When training a VAE, real-world training images are fed into the encoder to produce a vector. This vector is then sampled to provide input for the decoder, which attempts to recreate the original real-world image. After sufficient training, randomly generated vectors should produce new, hitherto unseen images. While VAEs are still in use today (e.g., the open-source video deepfake software FaceSwap),⁵⁵ GANs have become far more popular and influential. A look at Google Trends reveals that over the past five years, searches for "generative adversarial networks" have, on average, been 30% more



” Early generative models amazed the general public with their ability to synthesize ultra-realistic imitations of human faces, today’s models create realistic, high-quality audio and video impersonations.

frequent than searches for “variational autoencoder”.

GANs are trained by having generator and discriminator models compete with each other. The generator attempts to create fake images that can fool the discriminator. The discriminator is shown both real and fake (generated) images and tries to distinguish between them. This iterative competitive process, where the generator improves its outputs based on the discriminator’s predictions and the discriminator learns to discern which images are fake, is at the core of GAN-based architectures. The training process is complex, requiring developers to calibrate the balance between improvements in the generator and discriminator models. GANs have produced excellent results in generating realistic original images. For example, StyleGAN2,⁵⁶ which is publicly available on thispersondoesnotexist.com, generates hyper-realistic images of human faces. Due to the public availability of the project, it is reasonable to feel concerned about the possible misuse of these original, realistic images of faces that appear to be human. If a malicious actor were to use such

an image in a fake profile, methods such as reverse image search might mislead other people into believing the profile is genuine. Fortunately, known inconsistencies⁵⁷ introduced by the model facilitate the detection of these synthetic images.

The Diffusion Model is another very promising model for generating realistic human faces. Its sample quality has recently surpassed that of state-of-the-art GANs.⁵⁸ However, due to the architecture of the model, training is more computationally demanding than for GANs.

Yet another exciting and novel VAE-based approach to image generation is DALL-E, created by OpenAI.⁵⁹ This model demonstrates an ability to generate images based on semantic textual descriptions without fine-tuning. If not appropriately regulated, such models might be used for generating fake images to complement disinformation narratives. For example, the model can generate multiple images of the same scene with slight variations (e.g., from different angles) and could therefore theoretically be used to reinforce



the credibility of fake images. Around 250 million image-text pairs collected from the internet were used to train the DALL-E model with architecture similar to that of GPT-3. This is a computationally expensive model. Even if generating only low resolution (256x256) images, a GPU device with at least 24 GB of memory is required to house the pre-trained model.

Synthetic audio (voice)

Like images, deep learning algorithms make it possible to generate audio deepfakes for malicious purposes. Audio deepfakes are especially dangerous because they embed biometrics and can be used in speech-based identity verification systems.⁶⁰ Currently, the best audio deepfake methods work as text-to-speech (TTS) synthesis or voice conversion (VC) systems because the task of generating audio from scratch remains challenging.⁶¹

Most of the TTS models that are widely used today, such as Google Translator,⁶² have poor audio quality (low sample rate, intonation, etc.). However, quite a few sophisticated models are available, for example, Tacotron,⁶³ Wavenet⁶⁴ and DeepVoice3.⁶⁵ These models are trained on actual recordings to generate more natural-sounding speech. It is also possible to generate a realistic voice clone of a particular person. Recent academic studies^{66,67} demonstrates generation of realistic voice using audio samples. One problem with TTS is that detailed audio transcriptions with timestamps and non-

verbal noise annotations are needed to train a high-performance model. It takes a great deal of time to do the annotating, denoising, etc. required for these transcriptions, as demonstrated by the video deepfake study⁶⁸ in which researchers used twelve hours of speech recordings of a politician to train the Tacotron 2 model.⁶⁹

Voice Conversion alters a source speaker's audio waveform to sound like the target speaker's voice but leaves the linguistic content unchanged. This kind of speech-to-speech conversion offers more flexibility than TTS methods because source speakers can control their own intonation to influence the intonation in the result. Often, TTS models are altered to accomplish VC by changing the encoding part of the network to receive audio instead of text (e.g., NAUTILUS).⁷⁰ An important limitation of VC methods is that they usually require sizeable parallel source and target speech data. Nevertheless, Descript⁷¹ demonstrates that by exploiting large amounts of multi-speaker non-parallel data, it is possible to achieve good results with only a few seconds of parallel audio.⁷²

There are multiple online services for TTS and voice cloning such as Google Cloud TTS,⁷³ Amazon AWS Polly,⁷⁴ Baidu TTS,⁷⁵ Overdub,⁷⁶ and iSpeech.⁷⁷ However, the providers of these services attempt to complicate the service exploitation for malicious purposes via specific agreements, limitations, and countermeasures. For example, a TTS service might not work



with text-audio pairs that have merely been uploaded but require the target voice owner to read a specific text.

Like with other AI domains, there are multiple open-source code repositories freely available that could be used to create audio deepfakes. Mozilla TTS⁷⁸ is a repository with a large selection of high-performance TTS models. Although a high level of technical expertise and high-quality data is required to train such models properly.

The main goal of any speech synthesis is to accurately mimic the salient aspects of a genuine voice, such as expressiveness, roughness, breathiness, stress, emotion, etc. Deepfake audio can already model these features, so it is often difficult for humans to distinguish fakes from genuine speech.⁷⁹ This is especially true for low-quality channels such as phone calls where generative defects are similar to the usual noise and distortions. To make a fake even more convincing, the generated audio can be improved manually with audio editing software such as Audacity⁸⁰ or Descript.⁸¹ The most realistic video deepfakes to date have been achieved with voice actors (e.g., Tom Cruise deepfake⁸²), but this may soon change.

Synthetic video

Creating convincing synthetic (fake) videos is a challenging task. It is currently much more common among disinformation actors to use a “cheapfake”⁸³ —a sample of

real media that has been altered using one or more simple techniques (e.g., splicing, slowing it down) to suggest a narrative different from those portrayed in the actual footage. Such low-cost fakes can still have a significant effect. A popular example is the video of House Speaker Nancy Pelosi talking about Donald Trump that has been slowed down to make her appear intoxicated. This video caused waves of public outrage in 2019.

The generation of fake videos with learning-based methods has become more accessible due to popular open-source repositories such as FaceSwap⁸⁴ and DeepFaceLab.⁸⁵ Deepfake creators need several videos depicting the target person and, for best performance, a computer with a powerful GPU. Target videos are split into frames and then processed before training. Numerous pre-trained models are available to simplify the data preparation pipeline including face detection, segmentation and rotation models together with essential photo editing tools that can be used to automatically swap the original face with a generated one. The model is then trained using newly extracted images of the target’s face and used to manipulate unseen frames, which are manually corrected and adjusted to yield the best results.

One of the main hurdles to generating fake videos has been the amount of data required. Because of this, both legitimate research efforts and malicious campaigns have primarily focused on video of public



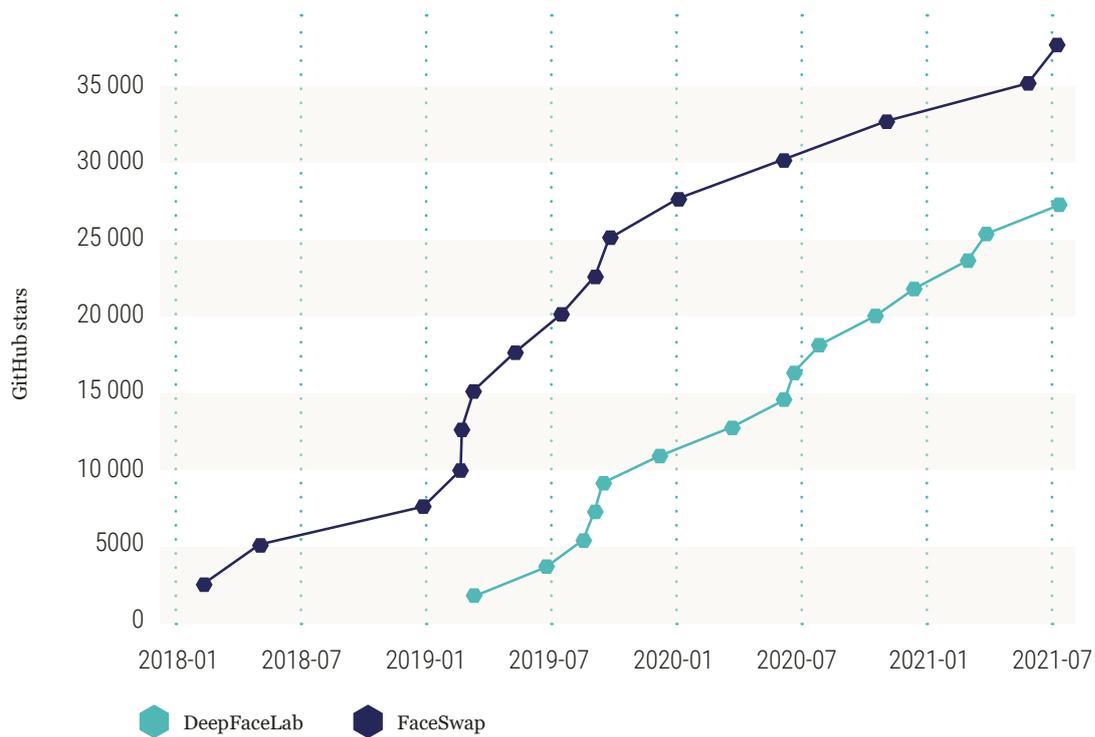


Figure 2. The popularity of the GitHub repositories DeepFaceLab¹²⁰ and FaceSwap¹²¹ is growing. The GitHub stars axis represents the number of platform user bookmarks of these repositories.

Graph created by the authors.

figures with a plethora of available data. However, Yao et al.⁸⁶ have developed an all-in-one tool for generating fake video with text-to-speech audio. Their tool can modify an existing video by changing the spoken text, facial expressions, and the intensity of gestures. The troubling aspect of this

excellent work is that the researchers could produce convincing talking head videos using as little as 30 seconds of training data. These recent developments mean that just about anyone can become a target of impersonation and other malicious activities.



Spreading disinformation

Bots, sock puppets, cyborgs

Fake accounts that employ varying levels of automation (sock puppets, cyborgs, bots) are used in influence operations on social media.^{87,88,89,90} This section discusses how nefarious actors use AI models and tools to create fake accounts, and to automate engagement. Automated accounts seem to be having a significant impact on public debates in social media. For example, in 2018, Twitter released a dataset containing 3,841 accounts affiliated with the Russian Internet Research Agency (IRA)⁹¹—probably the best-known organization perpetrating influence campaigns on the internet. Using social bots is an inexpensive way for malicious actors to spread large amounts of disinformation and influence a society’s perceived consensus around divisive issues.

Sock puppets

Sock puppet accounts are fake accounts that have been created to deceive, operated in mass by real people but dissociated from any real identity.⁹² Such accounts are used by actors with malign intent and by the investigators working to uncover and research information activities. In December of 2019 a network of over 900 pages, groups, and accounts on Facebook and Instagram was identified as a case of “coordinated

inauthentic behavior” and taken down. This was the first verified case of a generative adversarial network being used to create deepfake profile pictures at scale.⁹³ This type of fake image can still be detected by humans who can observe, for example, an “asymmetry in the glasses”, ears, hair, or a “poorly defined background”.

Social bots and cyborgs

In this study, we define *social bot* (or simply *bot*) as a fully automated online social network (OSN) account and *cyborg* as a partially automated bot account, which from time to time can be used by human to interact and disseminate specific narratives. Social bots and cyborgs have many practical pro-social applications, e.g., public service announcements, customer service, disaster information, etc. Twitter emphasizes that they target only automated accounts (bots) that are used for manipulation.⁹⁴ A typical information warfare application for bot networks is to generate artificial support or opposition to a political campaign, or to amplify false or biased stories to influence public opinion. This is known as *astroturfing*, alluding to the falsification of grassroots opinion. Such bots are often put into place and tasked with posting irrelevant information until the botmaster sends a command to engage with specific content



” Using social bots is an inexpensive way for malicious actors to spread large amounts of disinformation and influence a society’s perceived consensus around divisive issues

positively or negatively by, for example, liking, sharing particular political posts, or posting specific messages defined by the botmaster.

Prevalence of AI methods in bot networks

Botmasters are generally not interested in sharing their botnet codebases, so researchers have little opportunity to examine the source code of social media bots. Therefore, it is difficult to know precisely how widespread ML methods have become in bot applications. However, several techniques can be used to arrive at a working estimate.

Investigating open-source bot projects

One way to estimate the importance of AI methods in developing open-source bots is to analyze open-source code repositories. Studies^{95,96} performed in 2019 examined a total of 40 301 code repositories from GitHub (38 600), GitLab (1293), BitBucket (408), and SourceForge (25). The top four social media platforms for bot software

repositories were: Telegram ~20 000, Twitter ~10 000, Facebook ~3000, and Reddit ~3000. Bot software repositories of <1500 were also found on Skype, Instagram, YouTube, WhatsApp, Tumblr, VKontakte, Snapchat, and Pinterest. Researchers selected the 85 most active repositories for manual inspection and found that 22 were used to post predefined content and to like specific hashtags. Moreover, they found that 15 of the 85 repositories had chat-related functionality, and 14 repositories created new posts “based on former posts or an external text database.” Only 2 out of the 85 source code repositories were identified as “stand-alone” or in other words, ready to use operational bot scripts which could be implemented to mimic a human user. These open-source repositories mostly consisted of the components needed to build bots but did not maintain fully automated bot pipelines. Therefore, we can conclude that there is an additional development cost to assembling a bot network.⁹⁷

For more up-to-date information, in November 16, 2021 we found 13 922 repository



results for our “Twitter bot” query⁹⁸, but not all repositories had bot source codes. The use of Python language dominates as well as various popular deep learning and natural language processing libraries such as PyTorch, Tensorflow, nltk and transformers. Some bot developers use scraped tweets to train the GPT-2 language model to generate new human-like tweets.⁹⁹

Observations from bot markets

Another way to estimate the prevalence of AI tools in social bots is to investigate the availability and advertised features of bots in various markets. In 2019, a study inspected 30 bot markets on the surface web and 31 markets on the darknet.¹⁰⁰ The study found that “non-automated fake followers” (simulating only social support and digital connections) were relatively cheap—3 000 Instagram followers could be purchased for 18.24 EUR, i.e., roughly one cent per follower. In contrast, “active creation of content” was very rare and much more expensive—100 random Instagram comments were traded at 62 cents per comment on a Darknet crypto market.¹⁰¹ Current AI limitations might lead one to question the significance of AI in existing bot networks. The relative rarity and high cost of active, content-creating inauthentic accounts suggest that in 2019 it was challenging to develop fully automated content-creating botnets. Having to rely on manual labor for content preparation has restricted the operations of “troll farms”.

Disinformation enabled by social networks and data brokers

AI provides internet consumers with a wide range of intelligent, data-based services that are highly optimised and effective at delivering various products and services via ads to the target audience. However, despite their AI-enabled decision making, these complex systems cannot distinguish between malicious and general-purpose products, services, and information. This allows malicious actors to exploit AI capabilities indirectly by manipulating the engagement data. We discuss the manipulation of AI-enabled services in greater detail below.

Micro-targeted disinformation spread through existing online advertising infrastructure

Advertising platforms and social media sites can be used to spread disinformation by means of ads tailored to the interests of targeted vulnerable groups.¹⁰² Messages designed to micro-target individual users have a high chance of resonating, and the engaged users are unlikely to report such messages to website administrators.¹⁰³ Therefore, micro-targeting is generally an effective way to spread disinformation. One way to combat this type of exploitation is for those who maintain the advertising infrastructure to require greater transparency regarding the ad-content, including targeted political ads.



Content recommendation algorithms facilitate polarized echo chambers

Social media platforms are inherently susceptible to disinformation campaigns due to their dependence on advertising income and the way in which algorithms maximize user engagement.¹⁰⁴ Social media users tend to engage with information that supports their beliefs. The content recommendation algorithms used by YouTube and Facebook learn from user activity and so amplify the tendency to look only at a narrow range of content.¹⁰⁵ This creates a positive feedback loop in which users are progressively shown more of the same sort of content containing biased information that supports their existing ideas and does not challenge their beliefs. Such content bubbles are called “echo chambers”. Some echo chambers become politically polarized,¹⁰⁶ creating the perfect conditions for exploitation by divisive disinformation campaigns—radical content

gets distributed to users who are likely to agree with it and have their “sense of reality” reinforced by other users in the same echo chamber.

Data brokers

Data brokers are companies that collect, aggregate, and trade data for commercial gain.¹⁰⁷ Such companies accumulate large datasets that can be used for various kinds of user profiling (e.g., psychological, political, or commercial) and other analyses that can facilitate the creation of more precisely targeted disinformation campaigns. This was demonstrated in 2016 when Cambridge Analytica harvested up to 87 million Facebook user profiles and used the information to influence voters in the 2016 US Presidential elections.^{108,109} Some studies claim that the Cambridge Analytica story has inspired other organizations to use similar tactics.¹¹⁰



Possible future trends for AI-powered disinformation

A significant amount of AI research is *open source*, meaning that the software is freely available and may be both modified and redistributed. This spirit of openness leads to rapid development from research to application. If this trend continues, we can expect that once their capabilities exceed those of the current tools of disinformation, AI technologies will be rapidly adopted by malign actors. **Based on our research, we expect the following trends in AI-powered disinformation to be realized within the next five years:**

Optimised resource requirements for text generation

State-of-the-art NLP models can now generate highly realistic paragraph-length texts. The main limitation of these models is that they require vast computational resources and cannot be run on a personal computer. Large-scale model training requires even greater resources and is not practical for small or medium-sized companies and institutions. The development of more resource-efficient neural network architectures and training methods is an active field of research. Breakthroughs in efficiency could rapidly increase the use of high-quality NLP models to generate articles, posts, and comments in support of any narrative.

Improvements in machine translation

Machine translation makes it possible to reuse disinformation content in any number of languages. This method is currently

effective for major world languages with significant populations that have provided the volume of text necessary to train automatic translation models to work with reasonable accuracy. Incremental improvements in machine translation will continue, refining translation services to and from both major and minor languages and directly impacting the reach of reused disinformation content.

Major AI advances in the hands of state actors and large corporations and states

The most significant breakthroughs in AI research usually require massive datasets and thousands of GPUs or CPUs for implementation (e.g., GPT-3, AlphaGo, others). The cost of training such algorithms is in the order of millions or tens of millions of US dollars.¹¹¹ These large-scale models are optimised and downsized so they can be deployed on smaller hardware infrastructures. However, state-of-the-art models usually require significant hardware



infrastructure, access control over immense datasets, and significant human resources. AI research and development in technology-oriented but less democratic countries can evolve rapidly thanks to unlimited access to data collected on citizens.¹¹² Emerging AI-powered technology and 5G capability will enable data collection at a previously unseen scale.

Wider use of deepfakes

The AI research community contributes steadily to the development of open-source tools for deepfake generation, moving towards more accessible tools and easier generation of deepfakes. However, data preparation and processing continue to be a bottleneck. For example, creating a realistic audio deepfake from scratch with no prepared data of the target voice currently requires hundreds of hours of work by a human annotator.

An incremental increase in the quality of deepfakes

Generative modeling of images, video, audio is an active area of research, so it is reasonable to expect increasing fidelity in all types of deepfakes. However, the field could potentially come to a stagnant phase if fundamental improvements are needed in model performance to automatically synthesize ultra-realistic, temporally robust video deepfakes efficiently. Currently, models such as DALL-E can *imagine* (or generate an image of) only one scene, but

soon may be able to render a short video of a specific activity in the same way.

AI combining multiple domains

Another exciting research topic is the complementation of text produced by NLP models with images or video from the real world. In the future, a model such as OpenAI's DALL-E, which generates images based on descriptions in natural language,¹¹³ could be used to increase the persuasiveness of disinformation text by supporting it with generated images or video. Or vice versa, text could be generated to reinforce a narrative presented in visual or audio content. In this way, bots could automatically generate comments about posts with images or videos.

More general-purpose AI that does not require training and fine-tuning

It has recently been demonstrated that large AI models can be fine-tuned to accomplish new tasks without additional training. For example, after initial pretraining, OpenAI's CLIP can perform a wide variety of image analysis tasks (optical character recognition, action recognition in videos, geo-localization) without any training on the new data.¹¹⁴ GPT-3's ability to accomplish various unanticipated NLP tasks is another example. Such models require extensive training resources initially but can then be used to perform new tasks with considerably fewer resources. Using this type of model, the importance of AI in



disinformation content generation could increase.

Bots becoming better at mimicking humans

AI will also likely be used to efficiently imitate user activity patterns on social networks. The development of bots that can perfectly mimic a substantial percentage of user activity patterns could make botnets nearly undetectable. To keep up, social network administrators will have to implement much more rigorous security features and implement countermeasures such as more complex CAPTCHA tests for bots. Such developments would inconvenience users and likely result in a bad user experience of various services and interfaces. Moreover, cyborg activity would be encumbered but not stopped entirely.

Potential future risk—artificial general intelligence

The most exciting, and potentially most dangerous, development would be the creation of an artificial general intelligence (AGI). This is unlikely to happen within the next five years. The current level of development is far from that of a true AGI. In 2018, a sample of 352 AI experts believed on average that the development of a true AGI would require another 100 years of research (with ~75% probability of success).¹¹⁵ If successful, such a complex algorithm system would be capable of general reasoning at the level of an average human, posing the risk that the AGI could

be parallelized and scaled to create vast quantities of high-quality disinformation without being detected as artificial. Whether or not the creation of a true AGI is possible, AI researchers can shorten the time needed to reach something approximating AGI. This would require the development of an AI roadmap, a Manhattan Project-like undertaking.¹¹⁶



Conclusions

Currently, most AI models are trained and can operate in narrow domains only; these models are puppets mimicking human-like behavior or generating synthetic data with striking realism but still follow strict human-formulated rules. The recently introduced large-scale, more general-purpose models like GPT-3 are an exception. Practical considerations place fundamental limitations on the use of AI for disinformation, especially with regard to bots and long text generation, as social media platforms continue to improve their ability to detect malicious activity. This is a likely reason why currently **AI seems to be used only sporadically and has not been widely adopted for use in disinformation campaigns**. As their efficiency and performance increase, AI models are slowly being integrated into various content analysis and generation tools.

Deepfakes are becoming much more challenging to detect. Today, almost anyone can generate hyper-realistic images from scratch (e.g., human faces). Such images are already being used for disinformation (e.g., as fake account profile pictures). The detection of a single fake image or audio file is already difficult. State-of-the-art audio models can generate highly realistic human speech. Deception is especially likely to succeed when lower quality channels such as a phone call are used. Realistic video

deepfakes are the most complex and most time-consuming to generate as they require human supervision and manual modifications. However, one might question whether a successful disinformation campaign requires high-quality content. Perhaps the greatest potential harm lies in the speed and variety of disinformation content that can be generated or modified using AI-powered tools. Until now, disinformation operations have mostly made use of low quality deepfakes. This may have created a false sense of security, distracting us from the upcoming security risks this technology poses. An illusion of safety might keep us from recognising deepfakes of higher quality and thus cause us to accept them as genuine information entities.

The generation of fully automated high-quality content is challenging and impractical for non-experts and hobbyists.

Training a large-scale AI model from scratch to achieve state-of-the-art performance requires significant computational resources and training data that are often accessible only to large companies or institutions. Despite alarming demonstrations of AI capabilities, their practical applicability for the purposes of disinformation remains limited. An analysis of open-source projects and bot markets suggests that, at least for individual actors, social bots, even those



that rely on smaller content generation models, are not widely available. We also observed that only a minority of bot code projects provide stand-alone solutions—most projects involve partial automation. However, this situation may change in the near future due to significant improvements in algorithms and hardware efficiency or to wider accessible of AI tools for content generation.

Most AI applications for disinformation campaigns could be integrated into social bots, including content generation (text generation, deepfake profile pictures) and automatic bot control algorithms (reinforcement learning techniques,¹¹⁷ behavior optimised by genetic algorithms).¹¹⁸ The social bot domain is bound to benefit from AI techniques relevant to generative models (text, images, activity patterns). For example, current language models can generate paragraph-long texts that are already difficult to distinguish from text written by humans; the source code and pre-training parameters are available for some of these models. This means they can be used to generate text rapidly, and this capability can potentially be exploited for disinformation campaigns to create, for example, fake articles, posts, or comments in social media without modification and even to duplicate the content in multiple languages by means of machine translation.

Great states will be the most important disinformation actors and also the

likely AI-supervillains. Sophisticated disinformation actors most likely follow and use the ideas of the greater international AI research community in terms of fundamental scientific exploration. We foresee that such an AI-powered actor most likely will aim to 1) generate high-quality micro-targeted disinformation at scale, 2) use undetectable social botnets for the efficient spread of content, and 3) work on bot automation capabilities to increase the impact on societies. The automated generation of quality content at scale requires hired human operators. Currently, AI-based tools for content generation still require a significant amount of manual effort (e.g., prompt engineering, deepfake editing, audio data labeling). Undetectable social botnets also need human operator interventions and network infrastructure (e.g., many proxy servers). Large neural network models require extensive and costly infrastructure (e.g., ~12M\$ to train a GPT-3-sized text model). However, bot automation and content optimisation are spheres in which malign actors might aspire to gain a relative advantage in the long term.

Finally, one might perceive disinformation as an industry where private entities can tell a story better and faster than governments. If demand for disinformation exists (like the attempt to push the COVID anti-vax plot),¹¹⁹ the companies and influencers might capitalise on it, and thus their capacity to exploit AI as a disinformation tool might define the efficiency of their influence campaigns.



Literature

- 1 W. S. McCulloch and W. Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biology* Volume 52 № 1/2 (1990): 99–115.
- 2 O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. Berg, and L. Fei-Fei, 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision* Volume 115 (2015): 211–52.
- 3 A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks' in F. Pereira and C.J.C. Burges and L. Bottou and K.Q. Weinberger (eds) *Advances in Neural Information Processing Systems 25 (NIPS 2012): 26th Annual Conference on Neural Information Processing Systems 2012* (Morgan Kaufmann Publishers, Inc., 2012), pp. 1097–105.
- 4 Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification', *IEEE Conference on Computer Vision and Pattern Recognition* (2014): 1701–708.
- 5 D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G.V. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, 'Mastering the Game of Go with Deep Neural Networks and Tree Search', *Nature* Volume 529 (2016): 484–89.
- 6 A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, 'Improved Protein Structure Prediction Using Potentials from Deep Learning', *Nature* Volume 577 (2020): 706–10.
- 7 DeepMind blog by the AlphaFold team, *AlphaFold: A Solution to a 50-year-old Grand Challenge in Biology*, DeepMind website, 30 November 2020.
- 8 C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. W. Pachocki, M. Petrov, H.P. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, 'Dota 2 with Large Scale Deep Reinforcement Learning' arXiv, 13 December 2019.
- 9 O. Vinyals, I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, A. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps and D. Silver, 'Grandmaster Level in StarCraft II using Multi-agent Reinforcement Learning', *Nature* Volume 575 (2019): 350–54.
- 10 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, 'Language Models are Few-Shot Learners' in H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds), *Advances in Neural Information Processing Systems*, Volume 33, (2020):1877–901.
- 11 A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, 'Zero-Shot Text-to-Image Generation', arXiv, 26 February 2021.
- 12 K. Drobotowicz, M. Kauppinen and S. Kujala, 'Trustworthy AI Services in the Public Sector: What Are Citizens Saying About It?' in F. Dalpiaz and P. Spoletini (eds), *Requirements Engineering: Foundation for Software Quality*, 27th International Working Conference, REFSQ 2021, Essen, Germany, 12–15 April 2021, Proceedings, pp. 99–115.
- 13 J. Schwartz, 'Tagging Fake News on Facebook Doesn't Work, Study Says', Politico, 11 September 2017. [Accessed 16 August 2021]
- 14 J. Akers, G. Bansal, G. Cadamuro, C. Chen, Q. Chen, L. Lin, P. Mulcaire, R. Nandakumar, M. Rockett, L. Simko, J. Toman, T. Wu, E. Zeng, B. Zorn and F. Roesner, 'Technology-Enabled Disinformation: Summary, Lessons, and Recommendations', arXiv, 3 January 2019.
- 15 R. M. Chesney and D. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *California Law Review*, Volume 107 № 6 (December 2019).
- 16 M. Landon-Murray, E. Mujkic and B. Nussbaum, 'Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of Fake News, Social Media, and Artificial Intelligence', *Public Integrity*, Volume 21 № 5 (2019): 512–22.
- 17 Proposal for a Regulation [Laying Down Harmonised Rules on Artificial Intelligence \(Artificial Intelligence Act\) And Amending Certain Union Legislative Acts](#), Brussels, European Commission, 21 April 2021.
- 18 M. Caldwell, J. T. A. Andrews, T. Tanay and L. D. Griffin, 'AI-enabled Future Crime', *Crime Science* Volume 9 № 1 (2020):



- 1–13.
- 19 I. Deksnite, 'How AI Can Create and Detect Fake News', *Forbes*, 12 September 2019. [Accessed 23 July 2021]
 - 20 J. Wakefield, "'Dangerous' AI Offers to Write Fake News", *BBC News*, 27 August 2019 [Accessed 21 July 2021]
 - 21 From an online interview with Mr Viktoras Daukšas, Head of DebunkEU.org, conducted by the authors on 22 June 2021.
 - 22 JonasCZ, 'A Guide to Preventing Webscraping', GitHub repository. [Accessed 15 July 2021]
 - 23 F. H. Alqahtani and F. A. Alsulaiman, 'Is Image-based CAPTCHA Secure Against Attacks Based on Machine Learning?', *Computers & Security*, Volume 88 (2019) Article 101635.
 - 24 V. Carle, 'Web Scraping Using Machine Learning', Dissertation. KTH, School of Electrical Engineering and Computer Science (EECS). 2020.
 - 25 Advertisement for Next-Gen Residential Proxies With AI & ML Power, Oxylabs website. [Accessed 15 July 2021]
 - 26 Carle, 'Web Scraping Using Machine Learning'.
 - 27 J. Dzieza, 'Why CAPTCHAs Have Gotten So Difficult', *The Verge*, 1 February 2019. [Accessed 15 July 2021]
 - 28 Alqahtani & Alsulaiman, 'Is image-based CAPTCHA secure?'.
 - 29 A. Radford and K. Narasimhan, 'Improving Language Understanding by Generative Pre-Training', Pre-print, Open AI, 2018.
 - 30 J. Devlin, M. Chang, K. Lee and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *NAACL-HLT*, 2019.
 - 31 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and L. Sutskever, 'Language Models are Unsupervised Multitask Learners', *OpenAI blog*, 1(8), 9.
 - 32 N. Keskar, B. McCann, L. Varshney, C. Xiong and R. Socher, 'CTRL: A Conditional Transformer Language Model for Controllable Generation', *arXiv*, 2019.
 - 33 C. Li, X. Gao, Y. Li, X. Li, B. Peng, Y. Zhang and J. Gao, 'Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space', *EMNLP*, 2020.
 - 34 Brown et al., 'Language Models are Few-Shot Learners'.
 - 35 Ibid.
 - 36 L. Fei-Fei, R. Fergus, and P. Perona. 2006. [One-shot learning of object categories](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.
 - 37 B. Buchanan, A. Lohn, M. Musser and K. Sedova, Truth, Lies, and Automation How Language Models Could Change Disinformation (Center for Security and Emerging Technology, Walsh School of Foreign Service, Georgetown University) May 2021.
 - 38 K. McGuffie and A. Newhouse, 'The Radicalization Risks of GPT-3 and Advanced Neural Language Models', *arXiv*, 2020.
 - 39 Brown et al., 'Language Models are Few-Shot Learners'.
 - 40 Ibid.
 - 41 M. Hutson, 'Robo-writers: The Rise and Risks of Language-generating AI', *Nature*, Volume 591 (2021): 22–25.
 - 42 M. Chen, J. Tworek, H. Jun, Q. Yuan, H. Ponde, J. Kaplan and W. Zaremba, 'Evaluating Large Language Models Trained on Code', *arXiv*, 2021.
 - 43 S. Black, L. Gao, P. Wang, C. Leahy and S. Biderman, '[GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow](#)', *Zenodo*, 21 March 2021.
 - 44 A. Spadafora, '[China Outstrips GPT-3 with Even More Ambitious AI Language Model](#)', *TechRadar*, 4 June 2021.
 - 45 Sberbank AI/ru-gpts, 'Russian GPT3 models', GitHub repository. [Accessed 15 July 2021].
 - 46 K. L. Chiu and R. Alexander, 'Detecting Hate Speech with GPT-3', *arXiv*, 2021.
 - 47 D. Heaven, 'OpenAI's New Language Generator GPT-3 is Shockingly Good—and Completely Mindless', *MIT Technology Review*, 20 July 2020. [Accessed 16 August 2021].
 - 48 K. Johnson, 'The Efforts to Make Text-Based AI Less Racist and Terrible', *Wired*, 17 June 2021. [Accessed 16 August 2021].
 - 49 Hugging Face website. [Accessed 15 July 2021]
 - 50 C. Raffel, N.M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, '[Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#)', *arXiv*, 2020.
 - 51 NVIDIA Riva, *NVIDIA website*. [Accessed 16 November 2021].
 - 52 From an email interview with Ms Alyssa Kann, Research Assistant at the Atlantic Council's Digital Forensic Research Lab, conducted by the authors on 23 June 2021.
 - 53 D. P. Kingma and M. Welling, 'Auto-Encoding Variational Bayes' *arXiv*, 2014.
 - 54 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, 'Generative Adversarial Networks', *arXiv*, 2014
 - 55 deepfakes/faceswap, 'Faceswap: Deepfakes Software For All', GitHub repository. [Accessed 15 July 2021]
 - 56 T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, 'Analyzing and Improving the Image Quality of StyleGAN', *arXiv*, 2020.
 - 57 Deconstructing Deepfakes—How do they work and what are the risks? *US Government Accountability website*, [Accessed 16 November 2021].
 - 58 P. Dhariwal and A. Nichol, 'Diffusion Models Beat GANs on Image Synthesis', *arXiv*, 2021.
 - 59 A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, 'Zero-Shot Text-to-Image Generation', *arXiv*, 2021.
 - 60 Y. Gao, T. Vuong, M. Elyasi, G. Bharaj and R. Singh, 'Generalized Spoofing Detection Inspired from Audio



- Generation Artifacts', arXiv, 2021.
- 61 M. Masood, M. Nawaz, K. M. Malik, A. Javed and A. Irtaza, 'Deepfakes Generation and Detection: State-of-the-art, Open Challenges, Countermeasures, and Way Forward', arXiv, 2021.
- 62 Google Translate website. [Accessed 15 July 2021].
- 63 Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio Q. V. Le, Y. Agiomyriannakis, R. Clark and R. Saurous, 'Tacotron: Towards End-to-End Speech Synthesis', INTERSPEECH 2017 conference paper, arXiv, 2017.
- 64 A. V. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, 'WaveNet: A Generative Model for Raw Audio' demo at the 9th ISCA Speech Synthesis Workshop, arXiv, 2016.
- 65 W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, 'Deep Voice 3: 2000-Speaker Neural Text-to-Speech', arXiv, 2017.
- 66 S. Ö. Arik, J. Chen, K. Peng, W. Ping and Y. Zhou, 'Neural Voice Cloning with a Few Samples', 32nd Conference on Neural Information Processing Systems, Montréal, Canada, January 2018.
- 67 J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi and T. Kinnunen, '[Can We Steal Your Vocal Identity From the Internet?: Initial Investigation of Cloning Obama's Voice Using GAN, WaveNet, and Low-quality Found Data](#)', conference paper at Speaker Odyssey 2018: The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018.
- 68 T. Dobber, N. Metoui, D. Trilling, N. Helberger and C. de Vreese, 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?', *The International Journal of Press/Politics*, Volume 26 № 1 (2020): 69–91.
- 69 J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyriannakis and Y. Wu, 'Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions', submitted to the IEEE International Conference on Acoustics, Speech and Signal Processing (2018): 4779–783.
- 70 H. Luong and J. Yamagishi, 'NAUTILUS: a Versatile Voice Cloning System', submitted to the IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume 28 (2020): 2967–981.
- 71 Descript website. [Accessed 15 July 2021]
- 72 H. Luong and J. Yamagishi, '[Speech Samples for "NAUTILUS: a Versatile Voice Cloning System"](#)'.
- 73 Google Cloud, Text-to-Speech website. [Accessed 15 July 2021]
- 74 Amazon, Amazon Polly website. [Accessed 15 July 2021]
- 75 Baidu AI, Baidu TTS website. [Accessed 15 July 2021]
- 76 Descript, Overdub website. [Accessed 15 July 2021]
- 77 iSpeech website. [Accessed 15 July 2021]
- 78 Mozilla, Mozilla TTS website. [Accessed 15 July 2021]
- 79 Masood et al., 'Deepfakes Generation and Detection'.
- 80 Audacity website. [Accessed 15 July 2021]
- 81 Descript website. [Accessed 15 July 2021]
- 82 [Tome Cruise Deepfake](#), Youtube. [Accessed 21 November 2021]
- 83 B. Paris and J. Donovan, 'Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence', *Data & Society*, 18 September 2019.
- 84 deepfakes/faceswap, 'Faceswap: Deepfakes Software For All'.
- 85 I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, Mr. Dpfks, C. S. Facenheim, Luis RP, J. Jiang, S. Zhang, P. Wu, B. Zhou and W. Zhang, 'DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework', arXiv, 2021.
- 86 X. Yao, O. Fried, K. Fatahalian and M. Agrawala, 'Iterative Text-based Editing of Talking-heads Using Neural Retargeting', arXiv, 2020.
- 87 R. Fredheim and K. Van Sant, Robotrolling 2019/4, (Riga: NATO Strategic Communications Centre of Excellence, 2019).
- 88 R. Manokara and M. Paramonova, Manipulation Ecosystem of Social Messaging Platforms, (Riga: NATO Strategic Communications Centre of Excellence, 2020).
- 89 R. Fredheim, M. Paramonova, and K. Van Sant, Robotrolling 2020/1, (Riga: NATO Strategic Communications Centre of Excellence, 2020).
- 90 S. Bay, A. Dek, I. Dek, and R. Fredheim, Social Media Manipulation Report 2020, (Riga: NATO Strategic Communications Centre of Excellence, 2021).
- 91 C. Kriel and A. Pavliuc, 'Reverse Engineering Russian Internet Research Agency Tactics Through Network Analysis', *Defence Strategic Communication*, Volume 6, (Riga: NATO Strategic Communications Centre of Excellence, 2019), pp. 199–227.
- 92 G. Pisciotto, M. Somenzi, E. Barisani and G. Rossetti, 'Sockpuppet Detection: a Telegram Case Study', arXiv, 2021.
- 93 B. Nimmo, C. S. Eib, L. Tamora, K. Johnson, I. Smith, E. Buziashvili, A. Kann, K. Karan, E. P. de León Rosas and M. Rizzuto, '#OperationFFS: Fake Face Swarm', *Graphika*, 20 December 2019.
- 94 Y. Roth and N. Pickles, 'Bot or Not? The Facts about Platform Manipulation on Twitter', Twitter blog, 18 May 2020.
- 95 D. Assenmacher, L. Adam, L. Frischlich, H. Trautmann and C. Grimme, 'Openbots', arXiv, 2019.
- 96 D. Assenmacher, L. Adam, L. Frischlich, H. Trautmann and C. Grimme, 'Inside the Tool Set of Automation: Free Social Bot Code Revisited', in Grimme, Preuss, Takes and Waldherr (eds), *Disinformation in Open Online Media*, from the First



- Multidisciplinary International Symposium on Disinformation in Open Online Media, (Springer, 2019), pp. 101–14.
- 97 Assenmacher et al., 'Openbots'.
- 98 [Github search query](#), [Accessed 16 November 2021]
- 99 M. Woolf, 'How to Build a Twitter Text-Generating AI Bot With GPT-2', Max Woolf's Blog, 16 January 2020.
- 100 L. Frischlich, N. G. Mede and T. Quandt, 'The Markets of Manipulation: The Trading of Social Bots on Clearnet and Darknet Markets' in Grimme, Preuss, Takes and Waldherr (eds), *Disinformation in Open Online Media, from the First Multidisciplinary International Symposium on Disinformation in Open Online Media*, (Springer, 2019), pp. 89–100.
- 101 Ibid.
- 102 F. N. Ribeiro, K. Saha, M. Babaei, L. Henrique, J. Messias, F. Benevenuto, O. Goga, K. Gummadi and E. M Redmiles, 'On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook' submitted to the Conference on Fairness, Accountability, and Transparency, 2019, arXiv, preprint 2018.
- 103 Ibid.
- 104 C. Stöcker, 'How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation' in Grimme, Preuss, Takes and Waldherr (eds), *Disinformation in Open Online Media, from the First Multidisciplinary International Symposium on Disinformation in Open Online Media*, (Springer, 2019), pp. 129–49.
- 105 M. M. Maack, ' "YouTube Recommendations are Toxic," Says Dev Who Worked on the Algorithm', TNW News, 6 May 2021.
- 106 A. Bessi, F. Zollo, M.D. Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi and W. Quattrociocchi, 'Users Polarization on Facebook and YouTube', PLOS ONE, Volume 11 № 8, 23 August 2016.
- 107 H. Twetman and G. Bergmanis-Korats, *Data Brokers and Security*, (Riga: NATO Strategic Communications Centre of Excellence, 2021).
- 108 Ibid.
- 109 S. Hoffmann, E. Taylor, and S. Bradshaw, *The Market of Disinformation*, Oxford Information Labs, 2019.
- 110 Ibid.
- 111 K. Wiggers, 'OpenAI's Massive GPT-3 Model is Impressive, but Size Isn't Everything', VentureBeat, 1 June 2020.
- 112 C. Larson, 'China's Massive Investment in Artificial Intelligence has an Insidious Downside', Science, 8 February 2018. [Accessed 23 August 2021]
- 113 Ramesh et al., 'Zero-Shot Text-to-Image Generation'.
- 114 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, 'Learning Transferable Visual Models From Natural Language Supervision', arXiv, 2021.
- 115 K. Grace, J. Salvatier, A. Dafoe, B. Zhang and O. Evans, 'When Will AI Exceed Human Performance? Evidence From AI Experts', *Journal of Artificial Intelligence Research*, Volume 62 (2018): 729–54.
- 116 J. C. Levin and M. M. Maas, 'Roadmap to a Roadmap: How Could We Tell When AGI is a "Manhattan Project" Away?' submitted to the 1st International Workshop on Evaluating Progress in Artificial Intelligence-EPAI, August 2020.
- 117 A. Saleh, N. Jaques, A. Ghandeharioun, J. Shen and R. Picard, 'Hierarchical Reinforcement Learning for Open-domain Dialog', *AAAI-20 Technical Tracks 5*, Volume 34 № 05, (2020): 8741–48.
- 118 S. Cresci, M. Petrocchi, A. Spognardi and S. Tognazzi, 'Better Safe Than Sorry: An Adversarial Approach to Improve Social Bot Detection', *Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 47–56.
- 119 C. Haynes and F. Carmichael, 'The YouTubers who Blew the Whistle on an Anti-vax Plot', BBC News, 25 July 2021. [Accessed 25 July 2021]
- 120 [DeepFaceLab](#), GitHub repository. [Accessed 16 November 2021]
- 121 [FaceSwap](#), GitHub repository. [Accessed 16 November 2021]



