December 2021

AI and the Future of Disinformation Campaigns

Part 2: A Threat Model

CSET Policy Brief

AUTHORS
Katerina Sedova
Christine McNeill
Aurora Johnson
Aditi Joshi
Ido Wulkan

Executive Summary

The age of information enabled the age of disinformation. Powered by the speed and volume of the internet, disinformation has emerged as an instrument of strategic competition and domestic political warfare. It is used by both state and non-state actors to shape public opinion, sow chaos, and erode societal trust. Artificial intelligence (AI), specifically machine learning (ML), is poised to amplify disinformation campaigns—influence operations that involve covert efforts to intentionally spread false or misleading information.[1]

In this series, we offer a systematic examination of how AI/ML technologies could enhance these operations. Part 1 of the series described the stages and common techniques of disinformation campaigns.[2] In this paper, we examine how AI/ML technologies can enhance specific disinformation techniques and how these technologies may exacerbate current trends and shape future campaigns.

Our findings show that the use of AI in disinformation campaigns is not only plausible but already underway. Powered by computing, ML algorithms excel at harnessing data and finding patterns that are difficult for humans to observe. The data-rich environment of modern online existence creates a terrain ideally suited for ML techniques to precisely target individuals. Language generation capabilities and the tools that enable deepfakes are already capable of manufacturing viral disinformation at scale and empowering digital impersonation. The same technologies, paired with human operators, may soon enable social bots to mimic human online behavior and to troll humans with precisely tailored messages. These risks may be exacerbated by several trends: the blurring lines between foreign and domestic influence operations, the outsourcing of these operations to private companies that provide influence as a service, and the conflict over distinguishing harmful disinformation and protected speech.

We conclude that a future of AI-powered campaigns is likely inevitable. However, this future might not be altogether disruptive if societies act now. Mitigating and countering disinformation is a

whole-of-society effort, where governments, technology platforms, AI researchers, the media, and individual information consumers each bear responsibility.

Our key recommendations include:

Develop technical mitigations to inhibit and detect ML-powered disinformation campaigns. Social media companies and Congress should inhibit access to user data by threat actors and their proxies. The U.S. government and the private sector should increase transparency through interoperable standards for detection, forensics, and digital provenance of synthetic media. Chatbots should be labeled so that humans know when they are engaging with an AI system.

Develop an early warning system for disinformation campaigns. Expand cooperation and intelligence sharing between the federal government, industry partners, state and local governments, and likeminded democratic nations to develop a common operational picture and detect the use of novel ML-enabled techniques, enabling rapid response.

Build a networked collective defense across platforms. Online platforms are in the best position to discover and report on known campaigns. Because these campaigns may occur across multiple platforms it's important to share information quickly to enable coordinated responses. All platforms, regardless of size, should increase transparency and accountability by establishing policies and processes to discover, disrupt, and report on disinformation campaigns. Congress should remove impediments to sharing threat information while enabling counter-disinformation research. Platforms and researchers should formalize mechanisms for cross-platform collaboration and sharing threat information.

Examine and deter the use of services that enable disinformation campaigns. As ML-enabled content generation tools proliferate, they will be adopted by influence-as-a-service entities, further increasing the scale of AI-generated political discourse. Congress should examine the current use of these tools by firms providing

influence for hire. It should build norms to discourage their use by candidates for public office.

Integrate threat modeling and red-teaming processes to guard against abuse. Platforms and AI researchers should adapt cybersecurity best practices to disinformation operations, adopt them into the early stages of product design, and test potential mitigations prior to their release.

Build and apply ethical principles for the publication of AI research that can fuel disinformation campaigns. The AI research community should assume that disinformation operators will misuse their openly released research. They should develop a publication risk framework to guard against the misuse of their research and recommend mitigations.

Establish a process for the media to report on disinformation without amplifying it. Traditional media organizations should use threat modeling to examine how the flow of information to them can be exploited by disinformation actors and build processes to guard against unwittingly amplifying disinformation campaigns.

Reform recommender algorithms that have empowered current campaigns. Platforms should increase transparency and access to vetted researchers to audit and help understand how recommendation algorithms make decisions and can be manipulated by threat actors. They should invest in solutions to counter the creation of an information bubble effect that contributes to polarization.

Raise awareness and build public resilience against ML-enabled disinformation. The U.S. government, social media platforms, state and local governments, and civil society should develop school and adult education programs and arm frequently targeted communities with tools to discern ML-enabled disinformation techniques.

AI-enabled disinformation campaigns present a growing threat to the epistemic security of democratic societies. Our report focuses on the social media and online information environment because they will be primarily impacted by AI-enabled disinformation

operations. They are part of a larger challenge that has undermined societal trust in government and the information upon which democracies rely. While these recommendations may help stem the tide, the ultimate line of defense against automated disinformation is composed of discerning humans on the receiving end of the message. Efforts to help the public detect disinformation and the campaigns that spread it are critical to building resilience and undermining this threat.

Table of Contents

Introduction

> "The real problem of humanity is the following: we have paleolithic emotions, medieval institutions, and god-like technology."
>
> - Edward O. Wilson

On February 22, 2014, Ukrainians woke up to the news that their embattled president, Viktor Yanukovych, had fled to Russia. In the preceding months, Ukrainians had protested widespread corruption and Russia's pervasive influence. Russia viewed the protest and Yanukovych's flight as a clear failure of its "Ukraine policy."[3] Five days later, Russian special forces, operating without military insignia, fanned out through the Crimean Peninsula. Assisted by a local militia that they had cultivated for years, these forces seized the Crimean parliament, television and radio stations, and Ukrainian bases, perpetuating an illusion of a genuine rebellion.[4] Cyber operations cut off the means of communication with Kyiv and blockaded local Ukrainian military units.[5] In parallel with the physical operations, a sustained disinformation campaign bombarded the population of Crimea with narratives portraying the interim government in Kyiv as "fascists who threatened the Russian-speaking population with genocide."[6] The intense information onslaught on broadcast and social media was the beginning of a prolonged disinformation campaign effort to legitimize an occupation.[7]

Russia took the world by surprise by exercising a "new way of warfare for the 21st century war."[8] It encompassed the use of military, cyber, and influence operations and exploited societal fissures to disguise an invasion under the veil of self-determination. A month later, Russian special forces did it again. Masquerading as locals, they took over municipal administration buildings in Donets'k and Luhans'k and manufactured a rebellion, launching a Russia-Ukraine war that would claim nearly 14,000 lives, and that continues today. These tactics were accompanied by yet another disinformation campaign that exploited, weaponized, and

exacerbated societal polarization in Ukraine. In short, polarization became both a means and the ends of Russia's disinformation campaigns.

The age of information also enabled the age of disinformation. From its Cold War roots, disinformation has re-emerged as a tool of geopolitical strategic competition and domestic political warfare.[9] Powered by the speed and scale of the internet and leveraging digital marketing techniques, disinformation operations, which we define as the covert, intentional spread of false or misleading information, have weaponized social media platforms and fractured the information environment to sow discord and undermine trust.[10] Their full impact is difficult to measure empirically, yet they have certainly increased social discord and proliferated widely.[11]

The evolution of disinformation operations is underway. When Russia applied these tools to discredit domestic opposition and exploit fissures in democratic societies abroad, other nations took note.[12] Chinese and Iranian disinformation operations, also honed against domestic audiences and regional neighbors, are now deployed far beyond their borders to project power.[13] Today, 81 countries use social media to spread propaganda and disinformation, targeting foreign and domestic audiences through variety of automated and human-driven tactics.[14]

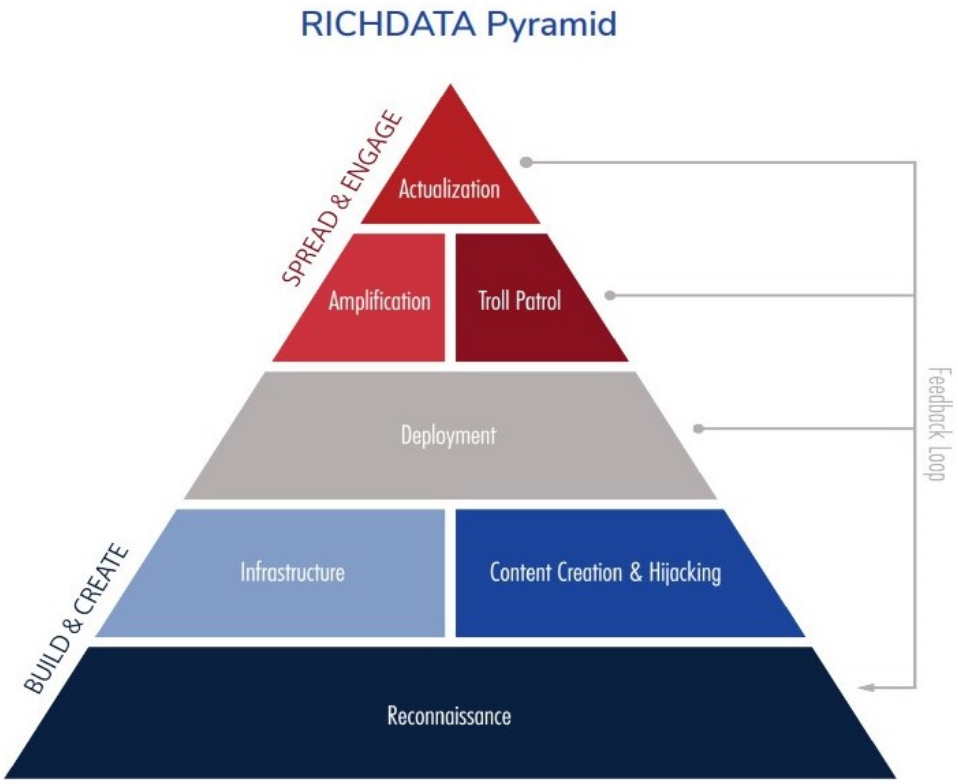
The techniques used in these campaigns highlight the technical, cognitive, and social nature of the disinformation problem. They capitalize on the central features of today's social media platforms—maximizing user engagement and connection. They overwhelm cognitive resources under stress from the intensity of modern-day information environment, exploiting biases and heightening emotions at the expense of rational decision-making.[15] They seek to deepen existing fissures within open societies, erode trust, and chip away at the shared values that form the common foundation critical to functioning democracies.

Artificial intelligence and its machine learning (ML) techniques have the potential to deepen the threat posed by disinformation campaigns from adversary nations and non-state actors. China and

Russia have stated the importance of winning the AI race and pledged investments ranging from hundreds of millions to billions of dollars in AI research and development.[16] The National Security Commission on Artificial Intelligence sounded the alarm about the potential of AI technologies to "increase the magnitude, precision, and persistence of adversarial information operations."[17] Concerns about AI-generated video and audio impersonations, known as "deepfakes," have garnered significant attention from researchers and policymakers.[18] However, there are other AI capabilities, such as generative language models, ML-enabled chatbots, and audience segmentation and sentiment analysis techniques that may be more impactful. The diffusion of generative language models and the commercialization of "AI-as-a-Service" presents a global challenge as societies grapple with the potential of AI-generated disinformation to threaten their epistemic security, decision-making, and trust.[19]

In this paper, we examine how advances in AI, specifically ML, are likely to augment disinformation campaigns. We leverage the RICHDATA framework that we described in detail in our companion report, "AI and the Future of Disinformation Campaigns: The RICHDATA Framework."[20] This framework describes the building blocks and stages of disinformation campaigns from the perspective of those who build them.

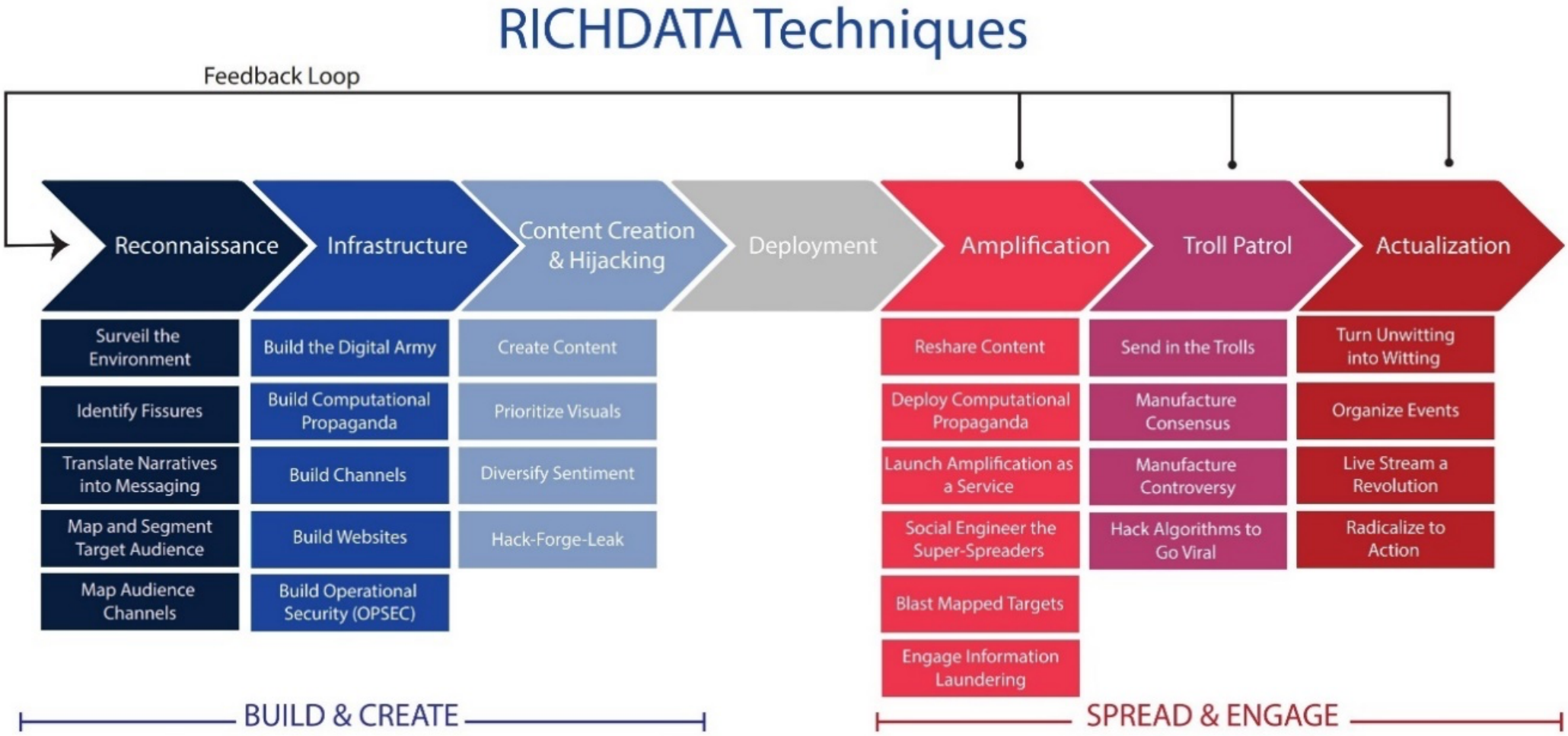
Figure 1. RICHDATA Pyramid



Source: CSET.

We break disinformation campaigns into multiple stages. Through **reconnaissance,** operators surveil the environment and understand the audience they are trying to manipulate. They require **infrastructure**—messengers, believable personas, social media accounts, and groups—to carry out their narratives. A ceaseless flow of **content**, from posts and long-reads to photos, memes, and videos, is a must to ensure their messages seed, root, and grow. Once **deployed** into the stream of the internet, these units of disinformation are **amplified by** bots, platform algorithms, and social-engineering techniques to spread the campaign's narratives. But blasting disinformation is not always enough: broad impact comes from sustained engagement with unwitting users through **trolling**—the disinformation equivalent of hand-to-hand combat. In its final stage, an influence operation is **actualized** by changing minds or even mobilizing witting or unwitting victims to action to sow chaos. Each of these stages has associated techniques that are depicted in the figure below.

Figure 2. RICHDATA Techniques



Source: CSET.

In this paper, we examine how advances in AI may enhance the stages and techniques of disinformation operations. We proceed in three parts. First, we identify the AI research areas most impactful for campaigns and how threat actors may apply them.

Second, we analyze salient features of current campaigns and the implications of future AI applications. Our research findings confirm that these technologies have the potential to augment disinformation operators and exacerbate the disinformation crisis through scale. Specifically, ML natural language processing algorithms are well suited to discerning emotion and sentiment from the data-rich online environment, opening up new avenues for social engineering. Large language models and the technology powering deepfakes, generative adversarial neural networks (GANs), provide a means to impersonate humans online and create disinformation at scale. These technologies may enable social bots to blend in by better mimicking human behavior patterns and enhance AI chatbots to target humans with precisely tailored messages. These risks are exacerbated by several current trends: the blurring lines between foreign and domestic disinformation, the outsourcing of disinformation creation to companies that provide influence services, the open culture of AI research, and the potential dual-use nature of many AI capabilities. All of the above may converge to further erode trust and accelerate the spiral of cynicism which creates fertile ground for further seeds of disinformation to take root.

Finally, we offer recommendations for how governments, social media platforms, media, and AI developers can prepare and respond to the malicious use of AI for disinformation campaigns. Notably, while our study examines applications of AI/ML to disinformation operations from a threat actor perspective, this technology can likewise empower defensive solutions. An assessment of defensive applications to mitigate and counter disinformation is not within the scope of this report and is left for future studies.

AI/ML Implications for Disinformation Campaigns

Machine learning—the ability of machines to learn from complex and voluminous data, identify patterns, and make inferences about decision rules—can both complement traditional automated techniques already employed in disinformation campaigns, and augment the campaigns in novel ways.[21] The transformative contribution of ML to any field depends on the presence of three elements: data, algorithms, and computing power. [22] Disinformation operations are no exception.

Non-ML Automation for Data Collection

Not all automated techniques rely on ML; nevertheless, they can be critical to collecting and harnessing the data that enables machines to learn. Malicious actors can collect relevant data by using open-source tools or by creating custom ones. For custom tools, they may use the application programming interface (API), a mechanism that many platforms enable to allow other software to "talk" to and interact with them more efficiently through code. These tools enable key tasks such as crawling networks to scrape publicly available data, and formatting or conditioning the data for additional processing and analysis. There is a vibrant market of freely available, open-source data scraping tools to collect data from social media platforms and download them into an easy-to-process format, such as a spreadsheet. Data scrapers can often circumvent the platform's efforts to limit access to user's posts. For example, Twint, a popular Twitter scraper available for free on GitHub, allows downloading of tweets, bypassing the limitations on quantity and rate of download, and the requirement to create a Twitter account imposed on developers that access this data through Twitter API.[23] While they may not provide the breadth of data available through the API, scrapers nevertheless allow malicious actors to sidestep platform vetting procedures and to access publicly facing user data, often used for nefarious purposes.[24]

Robust reconnaissance relies on harvesting and harnessing data about the target information environment. Ingesting vast volumes of data from diverse sources is not a trivial exercise and may

require dedicated software engineers and data scientists to build these systems in-house. First, building a central repository of data—a data lake—requires an environment that can collect data from different sources and in different formats. For example, data may come in simple spreadsheets of web links, databases of articles, archives of papers, or stores of images and video.[25] Second, data needs a place to go, such as cloud storage from commercial cloud providers. Third, this data may need formatting before it is ready for data analytics additional processing. This process—sometimes called data wrangling—can be simple or complex, depending on sources of data.[26] Alternatively, a well-resourced actor may procure commercially available solutions from niche companies who provide pre-processing services for "big data" or end-to-end solutions from ingestion to insight.[27] Once the data is ingested, sanitized, and curated, operatives can move on to finding insights, an area where machine learning techniques excel.

Foundational Machine Learning Technologies

The key ML technologies for disinformation campaigns include natural language processing and generation, and GANs.

Natural language processing (NLP) and **generation (NLG)** allow ML systems to read, write, and interpret text. NLP combines computational linguistics, modeling of human language based on rules, and artificial neural networks.[28] While earlier versions of NLP applications were often rules-based systems explicitly programmed to perform specific tasks, today's NLP neural networks automatically extract, classify, and label elements of text, and assign statistical likelihood to each possible element.[29]

As of September 2021, the two most powerful language generation systems are Generative Pre-Trained Transformer 3 (GPT-3), built by OpenAI, and its Chinese language equivalent, the PanGu-Alpha, recently introduced by Huawei.[30] These companies trained large language models with 175 billion and 200 billion parameters, respectively, on a trillion words of human text through massive neural networks.[31] These systems work with humans: operators pose questions, provide instructions, or prompt with examples of writing, to which the system responds with text in the

same style. GPT-3 completes the text the operators provide by probabilistically choosing each next word from a series of options based on the patterns it learned by analyzing human writing.[32] Researchers are working with GPT-3 to create poetry, write op-eds, develop computer code, and even generate images.[33] Developers are incorporating GPT-3 into apps to improve chat, translation, and ads.[34] Diffusion of large language models is underway. A mere year since GPT-3 proved novel capability, more generative language models—some of them larger and more complex—have been announced or released from U.S., Korean, Russian, Israeli, and French companies, enhancing language generation in more languages.[35]

Generative Adversarial Networks (GANs) use two neural networks—a generator and a discriminator—to compete against one another. The generator attempts to create data to fool the discriminator. The discriminator seeks to detect fake data from the generator. As the neural networks compete, they learn from each other, and the competitive process produces an increasingly realistic output in image, video, and audio formats.

Disinformation actors can harness the latest AI/ML innovations cheaply by leveraging **transfer learning**, a technique that allows the knowledge gained by an ML model in one setting to apply to a different related setting.[36] Well-funded AI researchers spend significant time and computing power training large neural networks on large datasets to solve a novel task. Researchers often release these pre-trained models into open-source code repositories so other researchers can reproduce, learn, and build on the research. Anyone can download the released pretrained model and retrain on a smaller dataset to accomplish narrower tasks. This technique enables researchers to fine-tune algorithms to learn new tasks in a way that is faster and cheaper, and that uses less computing power and training data.[37]

Private firms specializing in disinformation for hire— "influence as a service"—are likely to augment their non-ML offerings with ML-enabled techniques. In a recent threat report, Facebook researchers highlighted the significant contribution to disinformation operations of public relations and digital marketing firms that specialize in

online manipulation using deceptive means.[38] Some firms already sell content generation services and amplification bots that follow, retweet, and like tweets, though it is unclear if these services are ML-enabled. Many of them have hundreds of thousands of international customers, including celebrities, media pundits, candidates for public office, and state-run influence outlets, such as China's news agency Xinhua.[39] Disinformation operators increasingly outsource some or all stages of building a campaign to such firms, receiving plausible deniability in return.

The AI research community can also be a source of new tools if they release their research without built-in mitigations against misuse. For example, well-resourced firms are likely to incorporate the fine-tuned ML systems into their menu of services and offer ML-enabled disinformation as a service. Disinformation actors are likely to draw on these services.

These ML technologies form a powerful foundation of techniques that can enhance disinformation operators' workflow. In the following sections, we discuss their demonstrated and potential use for building disinformation campaigns. For each section below we highlight in blue the stages of the RICHDATA framework that are most impacted by these advances.

Enhance Surveilling the Information Environment

ML-powered **predictive analytics** capabilities can help threat actors identify current and future social fissures. For example, economists at Google used the publicly available search trends databases "Google Trends" and "Google Correlate" to create an accurate prediction model for economic "nowcasting"—predictions of the very near-future economic indicators.[40] The same methodology could theoretically be leveraged with different data points from these databases to create an engine that can accurately understand and predict near-future social trends in specific geographic regions, though this is speculative.

Virality, the tendency of content or topics to circulate rapidly and widely online, is a key goal for threat actors to achieve maximum engagement.[41] Researchers have used ML techniques to accurately predict which news events are likely to go viral early on in the contagion process.[42] The ability to predict the virality of organic content can confer an information advantage to threat actors because it gives them more time to spin narratives in their favor and exploit organic content more effectively. It may also give them insights into which narratives are more compelling and may deserve greater resources. Predicting message virality is relevant in both initial reconnaissance and for amplification.

Advances in ML are poised to make tools for **social listening**—monitoring social media mentions of a topic—more powerful and more precise. Commercial social listening tools combine network crawling and social media monitoring for engagement and mentions, two key indicators of audience interest, with analysis of users' sentiments.[43] Some tools do not use ML, but simply allow a user to build a keyword search string and receive an alert when this term is encountered on the web, not including social media.[44] More sophisticated tools claim to listen to millions of conversations across social media platforms and news sites, gauge the sentiment of the audience, analyze images, and produce comprehensive analytics of conversations.[45] Some commercially available tools provide a glimpse into how ML is enabling powerful social listening with advanced multilingual text analysis and visual analytics.[46] Threat actors can use similar technologies to better understand the target information environment and to provide feedback throughout their campaigns. While these and similar tools have limitations, they will continue to evolve and improve with the application of ML capabilities.

Boost Identifying Fissures with Sentiment Analysis

Sentiment analysis, a research field within NLP, extracts subjective qualities—attitudes and emotions such as sarcasm, confusion, and suspicion—from text and interprets the text as negative, positive, or neutral. Some ML models can determine the degree of

polarization from user comments and posts with 94-96 percent accuracy.[47] In one study, NLP and text classification techniques were used to pick up on linguistic patterns and identify depression in Reddit users.[48] While some skepticism persists about the efficacy of its early versions, sentiment analysis continues to improve. More sophisticated ML-enabled tools trained on datasets curated by human linguists are better able to grasp the colloquial language, abbreviations, emoticons, emojis, and slang used on social media.

Sentiment analysis tools allow tech-savvy disinformation actors to identify targets who post supportively or critically about certain topics. The current heuristics for doing so depend on inferring information, such as political leanings, from other contextual data, such as groups users belong to and people they follow. Sentiment analysis could allow disinformation operators to bypass this and identify a user's leanings based on their posts. State-of-the-art offerings can extract sentiment from text in over 13 languages and deploy the capability on the disinformation operator's network.[49] While these commercial solutions can be expensive, new tools are appearing on the underground market and drawing on openly released research.[50]

Sentiment analysis could also help determine the alignment of groups or individuals based on characteristics such as race, gender, or tone. Threat actors could build personas that are likely to resonate with an individual or group, feeding into their existing beliefs and precisely targeting vulnerable audiences to drive up engagement and influence behavior.

Stance detection takes sentiment analysis a step further, using ML techniques to classify text against a predefined concept.[51] Stance detection algorithms can identify not only whether the sentiment of a post is negative or positive, but whether it agrees with a broader idea—such as atheism or feminism.[52] Operators could use this to identify supporters and their communities during the reconnaissance phase of an operation. While the use of these algorithms currently requires an understanding of supervised ML in addition to basic web scraping, they are likely to be integrated soon into publicly available tools.

Researchers have adapted and tested a stance detection system across English, French, Italian, Spanish, and Catalan, including exploring the content of a tweet and deciphering contextual information.[53] Stance detection enables more accurate targeting of individuals by providing a deeper understanding of a user's positions and the strength of their beliefs. Such targeting can be further refined by combining this information with user data available from data brokers and a user's public posts.

Enhance Segmenting Target Audiences

Similar to social listening, social **network analysis** tools help visualize social and knowledge networks. Some tools use network graph theory, statistical analysis, and visualization techniques to show how nodes—people or organizations—are linked together, and the nature of their relationships.[54] Others apply natural language capabilities to process unstructured data and identify the targets, groups, and key connections between them.[55] These or similar tools allow operators to identify connections within and between groups, and key players to target.

ML techniques can help infer the attributes of users by examining salient aspects of their network. In a 2019 study on the popular Chinese social media site Renren, researchers used available data, such as a user's friends list, to infer unstated information, such as their age, hobbies, and university majors.[56] This network analysis technique allows malicious actors to make predictions about groups of users to infer strengths of relationships, similarity in interests, and potential susceptibility to influence on topics based on the posts of those in one's network.

Threat actors may use tools provided by the platforms to augment their audience segmentation efforts. The ever-increasing demand for precision advertising drives platforms to give third party applications and organizations access to nuanced audience segmentation tools. Using available APIs, applications can interact with the platform in real time, collect data, analyze it, and post content. Some products use this legitimately acquired data along

with ML to predict consumer trends.[57] Other commercially available tools purport to use unsupervised ML techniques to segment audiences, tease out sentiment, and map personas, producing visualizations of consumer attitudes.[58] However, operators can use these same tools for malicious purposes.

Platforms connect advertisers with their target audiences of digital "look-alikes." ML algorithms can recognize when users belong to a particular target audience and when their pattern of behavior on the platform—such as the type of content they engage with—matches that of other users. Some platforms provide built-in ML-enabled services to help advertisers and political consulting firms find new audiences through matching techniques.[59] They collect data points about each of their users, including data from other websites they visit, apps on their phones, and location data, although privacy-savvy users can opt out of this collection.[60] If threat actors are able to masquerade as legitimate advertisers—particularly by hiring firms that provide influence services—they too can find new look-alike audiences to target based on users' past pattern of online behavior on and off the social media platforms.

Psychographics is an advanced form of demographics that includes profiling of values, interests, activities, and opinions of population segments for advertising and election targeting.[61] Proprietary technology from firms like now-defunct Cambridge Analytica built psychographic profiles of people on the basis of the Big Five Model of personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism.[62] While Cambridge Analytica harvested user data in violation of Facebook's terms of use and its claims of efficacy are disputed, there are other, legitimate ways to collect psychographic information, including surveys, questionnaires, quizzes, website analytics, browsing data, psycholinguistic dictionaries, and social media engagement (likes, clicks, posts). Some commercial solutions combine psychographics, sentiment analysis, and deep learning to extract insight from text, including sentiment, emotion, and personality traits.[63]

Combined with demographic and "pattern of life" data, psychographic profiling and similar publicly available tools can help

threat actors target their messages precisely, provided they have enough user data. Between voter files, political data firms, commercial data brokers, and social media platform data, candidates for public office can have access to over three thousand data points on every voter.[64] Some threat actors may also have access to this data. More than 3.5 billion users saw their personally identifiable information, emails, photos, and credit card data stolen in the top breaches of the twenty-first century, including biometric fingerprint data of 5.6 million U.S. government employees with security clearances.[65]

Build a Digital GAN Army

As GANs—the ML capability behind deepfakes—supercharge the production of believable images of nonexistent humans, threat actors have taken notice.[66] The use of ML-generated profile pictures by disinformation campaigns is accelerating. One website, thispersondoesnotexist.com, illustrates this phenomenon. With every refresh of the site, a new fake face emerges, with an image of exceptional quality whose inauthenticity is practically undetectable with the human eye. Days after the website was created, it came to the attention of malicious actors, who promptly incorporated the new face generation capability into their operations. Just weeks later, Facebook discovered high-quality ML-generated profile avatars in a disinformation campaign on the platform.[67] Facebook quickly banned the use of GAN-generated content and relied on other markers, such as coordination, to detect the campaign. GAN-generated photos remain difficult to detect in real time, as detection and generation technologies continue to leapfrog one another.

Diffusion of GAN-generated Avatars

In just one year, the use of GANs went from a rarity to a staple of disinformation campaigns. In June 2019, in the first publicly known case, Katie Jones, an inauthentic persona on LinkedIn with a GAN-generated avatar, connected with scholars and high-profile figures in the national security community in a likely espionage operation.[68] By December 2019, Facebook had removed an inauthentic network with GAN-generated avatars operating groups for a content provider connected with the Epoch Media Group, a hyper-partisan U.S. media outlet with a special emphasis on coverage of China.[69] In August of 2020, the "Spamouflage Dragon" operation deployed a network of GAN-generated avatars on Twitter and YouTube posting in support of the Chinese Communist Party and targeting U.S. audiences with messages critical of the U.S. response to the COVID-19 pandemic and its China policy.[70] Later in August, in a "Peace.Data" operation, a network of Russian Internet Research Agency (IRA)-linked personas with GAN-generated avatars on Facebook, Twitter, and LinkedIn, targeted left-wing audiences in the United States and the United Kingdom. Masquerading as editors for a media site, they hired unwitting freelance writers in both countries to write stories aimed at steering progressive voters away from moderate candidates.[71] By October 2020, a parallel "NAEBC" IRA-linked operation used a network of GAN-generated avatars to target right-wing audiences on mainstream platforms as well as Parler and GAB.[72] By early 2021, fake personas with GAN-generated avatars engaged in rhetorical combat in information environments across the Central African Republic, Mali, and Libya, and attacked the Belgian government's plans to limit access of "high-risk" suppliers to its 5G network.[73]

The barriers to accessing GAN-generated avatars are low. A novice operator can navigate to a site like thispersondoesnotexist.com and save the generated photo, yielding a unique profile picture. An actor with limited coding skills can automate navigating to the site

and scraping photos, yielding a library of untraceable avatars faster and at scale. Using these readily available tools, an operator can easily and cheaply build a campaign around a set of randomly generated avatars.

Advanced actors seeking to target a specific racial or ethnic group can get a tailored set of avatars with some ML effort. They can finetune pre-trained models for face generation such as StyleGAN2 on a targeted dataset of photos of individuals or groups. At the 2020 BlackHat cybersecurity conference, threat intelligence researchers demonstrated exactly this, warning about how easily threat actors can leverage publicly available ML models for deception.[74] Using a dataset of publicly available photos of Tom Hanks, they finetuned StyleGAN2 to generate new fake photos of the actor. If threat actors want to create a campaign targeting seniors in retirement homes or military veterans, they can assemble a dataset of publicly available photo samples, finetune a model on this data, and get custom versions of thisseniordoesnotexist or thisveterandoesnotexist. Armed with avatars, they can build a supply of nonexistent humans for a tailored campaign targeting a specific group.

In addition, generating faces to precise specifications on demand may soon be part of publicly available applications. Researchers have already built an AttentionalGAN model, capable of generating images from a text prompt. This technology is rapidly improving and may soon be able to generate a photo of a fake human from a textual description—a photo of a nonexistent person fitting specified parameters, generated on demand.[75] In another effort, developers have combined the GPT-3 language generation system with a StyleGAN2 application to do just that. A text prompt asks GPT-3 to produce "a white female with blonde hair and green eyes." Eight photos of nonexistent humans fitting the description emerge.[76] The developers of this experiment plan to release this software in a public beta, for marketing professionals to generate faces for advertising purposes. It is unclear if or how the creators plan to guard against its misuse. Access to such an application will make it trivial for disinformation operators to order human faces of a specific race and ethnicity, outfitting digital troops to act as credible messengers in a campaign targeting ethnic minorities.

Beyond avatars, fine-tuning a pre-trained model can be useful in other aspects of creating inauthentic persona profiles. The persuasiveness and longevity of a fake persona depends on an authentic look, an identity with presence across multiple social media platforms, and a credible back story.[77] Here too, a series of fine-tuning projects have created a pipeline of free images to build out the rest of an inauthentic profile. For example, leveraging openly available demos, threat actors can build a profile of someone who has a fake cat, posts photos of fake food, lives in a fake apartment, appreciates fake art, and has a fake LinkedIn resume working for a fake start-up.[78] Researchers and developers have begun cataloguing ongoing experiments with pre-trained models.* At best, these efforts offer a "go to" catalogue for malicious actors to manufacture fake lives. At worst, they demonstrate how disinformation operators or third-party "influence-as-a-service" firms can build custom semi-automated capabilities.[79]

Scale Up Content Creation

ML may be the biggest game-changer in the content creation stage of disinformation operations, with the potential to increase the scale and virality of campaigns. Recent breakthroughs in NLP coupled with synthetically generated text, memes, video, and audio could make content richer and more compelling. Such tools will increase content creation and automate substantial parts of this human labor-intensive process as part of the human-machine team.[80] This is not limited to text generation, as there are also prototypes to create realistic, narrated videos and images from text prompts. [81]

ML-powered language translation and autocomplete functions can help foreign disinformation actors sound authentic, eliminating awkward phrasing and grammatical mistakes. Here, even relatively basic software can pay dividends. Google Docs has integrated ML into its Smart Compose feature, identifying grammatical errors and

* See a catalogue of such experiments on http://thisxdoesnotexist.com.

autocompleting sentences in real time.[82] Hijacking authentic posts or hiring native English speakers to write content, as has occurred in previous campaigns, is noisy and increases the potential for human leaks or exposure, whereas predictive composition is fast and cheap.[83]

Large generative models offer threat actors an "autocomplete on steroids" capability.[84] In a disinformation content farm, large language generation systems like PanGu-Alpha or the openly available GPT-J could streamline the labor-intensive process of creating text content.[85] A few human operators working with a language generation system can create pipelines of blogs, short posts, memes, and "junk news" from a few prompts. These systems do not have to fully replace humans to be powerful additions. They can do the heavy lifting, producing posts with prescribed number of characters, keywords, and themes of the day.[86]

These systems can also lessen the cognitive load of human operators such as those employed in a content generation "farm" by the Russian IRA.[87] This could reduce the problem of overworked human operators reusing talking points, lines, and phrases because they have to meet demanding content production quotas over long shifts.[88] Future campaigns may be able to significantly scale down the number of human content creators by using language generation models, and minutes later get back an array of content tailored to an audience's political preferences and biases. These systems can allow humans to shape the output and experiment. Such systems never tire nor suffer from poor morale or a crisis of conscience.

The quality of disinformation narratives created by systems like GPT-3 is largely untested, but the machine's ability to replicate the style and viewpoints of prompts shows promise. With just a few samples of extremist content, researchers have demonstrated its ability to mimic the style of far-right extremist writing and produce manifestos from multiple viewpoints.[89] A recent CSET study indicates that GPT-3 can craft stories that fit a viewpoint, manipulate narratives with a slant, seed new conspiracy narratives, and draft divisive posts that exploit political wedges and advance

racial and ethnic stereotypes.[90] Experiments also show that its content is persuasive and appears authentic. For example, in one study, after seeing five tweet-sized messages written by GPT-3 and curated by humans, the percentage of survey responders opposed to sanctions on China doubled. In another study, 73 percent of human responders judged stories generated by GTP-3's predecessor model as credible and mostly indistinguishable from those written by journalists, particularly as the partisanship of content increased.[91] Skeptics argue that GPT-3 often gets the facts wrong and loses the logical structure of its arguments after a few paragraphs. However, inaccuracy and inconsistency are not necessarily impediments for disinformation campaigns, especially with human oversight to curate the machine's outputs.

More troubling yet is the impact that synthetic text generation systems may have on (Hack)-Forge-Leak operations.[92] Because generative language systems can mimic a style of writing from just a few samples, these capabilities could enable targeted and precise content production. Previous disinformation operations illustrated how threat actors have sought to weaponize forgeries of official correspondence to sow discord among U.S. allies.[93] Threat actors seeking to impersonate and discredit specific public figures or organizations could conceivably do so at scale by creating believable forgeries, infusing them into a payload of hacked emails, and releasing them. Researchers have already demonstrated GPT-3's ability to generate emails from bullet points or re-write them to sound more polite.[94] Malicious actors could post forged emails, looking stylistically authentic and without any metadata that could provide forensic insight, which could be leaked to a media outlet in the closing weeks of an election campaign, for example. Yet with no technical way to detect machine-generated text, media outlets would face the task of verifying the unverifiable. For a threat actor, there may not even be a need to hack. They can prompt a language model with a few public samples of a public figure's writing, forge thousands of believable emails or messages, and claim that the emails came from a hack, giving the dump the illusion of authenticity.

Thus far, there have been no publicly known cases of threat actors operationalizing large language models for disinformation

campaigns. There are also no effective methods to detect their use. To social media platforms and internet browsers, synthetic text looks like any text on the internet. This means that if these systems were deployed today, the platforms would likely not be able to distinguish between human and machine-generated text. Early on in its fielding, GPT-3's creators at OpenAI were concerned with the misuse of this technology. They decided to control and vet access to the system—a somewhat controversial decision—to enable innovation while guarding against misuse. But the proof of concept is out there.

What sets GPT-3 apart for now is the size of its training data and its neural network but the gap is closing fast. While recreating a model like this from scratch is expensive, these are no barriers for advanced actors. This is evidenced by China's Huawei announcing the PanGu-Alpha, a predominantly Chinese-language model, Russia's Sberbank producing a smaller version of GPT-3 Russian-language model, and South Korea's Naver releasing a Korean-language HyperClova generative language model.[95] The system also does not need to be as advanced to be effective. The chances of diffusion increase as GPT-3 is commercialized, as is the likelihood that large language models will continue appearing in the open source or offered on subscription basis as a service.[96] EleutherAI, a collective of AI researchers advocating for open publication of language models, openly released GPT-J, a 6 billion parameter model, and plans to release larger replicas of GPT-3. Israeli firm AI21 recently announced a public beta for its Jurassic 1 Jumbo system, a 178 billion parameter generative language model, and plans to provide access to its capabilities to wide array of organizations and individuals for a fee.[97] As open-source and subscription-based language models emerge, well-resourced disinformation actors are likely to take advantage of them to saturate the information environment with believable falsehoods.

Rapid innovation is also supercharging production of engaging visual content, potentially making the work of disinformation operators easier and faster. For example, Chinese tech giant Baidu created a prototype that generates more than one thousand narrated video summaries of news stories daily. Powered by computer vision and NLP, the system —VidPress—builds two-

minute videos in less time than it would take humans. It reads an article, synthesizes key ideas, weaves them into a script, converts text to audio, collects related images and video clips, and combines them into a short video to deliver the content in a matter of minutes.[98] While this prototype was built for news curation, it has direct applications for disinformation campaigns.

Text-to-Image conversion technology may soon enable the creation of images from a text description.[99] DALL-E, a neural network trained to create images from text captions expressible in natural language is the newest innovation from the OpenAI Lab.[100] DALL-E is a 12 billion parameter version of its sister model, GPT-3, and is able to generate images from text descriptions, including animals, objects, historic relics, and cityscapes. While its creations still lack believability, this area of research is rapidly growing. Researchers from the Beijing Academy of Artificial Intelligence recently unveiled Wu Dao 2.0, a 1.75 trillion parameter multimodal system that can generate multiple types of digital media such as images and text from a prompt to illustrate a concept.[101] Once the messages of the campaign are clear, images may eventually be created on demand to support the narrative. That said, it remains to be seen how well this technology can compare to the current visual disinformation, such as cartoons, caricatures, memes, and photoshopping adversary heads of state into unrelated contexts.

Even the province of humorous meme creation is no longer strictly human-only. Humans are still better at making the memes funny, but how long this will remain the case is unknown.[102] While a predecessor to GTP-3 trained on a dataset of jokes could mimic the written shape and style of some jokes, most lacked the quality of true humor.[103] However, humor is not a prerequisite for a meme to go viral—divisive and hateful memes spread too.

Exploit Synthetic Image, Video, and Audio – the Deepfake Problem

Deepfakes are a product of GANs, the same deep learning concept as the one used for generating synthetic avatars. In recent years,

the number of deepfakes circulating on the internet has grown exponentially. Their uses include the spreading of nonconsensual, fake intimate imagery, the impersonation of employees to defraud a company, and the taunting of opponents in political campaigns.[104] While highly anticipated, no deepfake videos appeared as part of foreign influence campaigns in the 2020 U.S. elections. However, the FBI continues to warn about the increased risk of their use in cyber and foreign influence operations.[105]

Making deepfakes is getting easier, and access to large quantities of data is no longer a barrier. In May 2019, researchers at the Samsung AI Lab in Moscow built animated, talking head videos of Albert Einstein and Leonardo da Vinci's *Mona Lisa* using a single image as input.[106] This was a significant breakthrough given that previous examples had required a large dataset of images of the specific individual.[107] Head-swapping, or juxtaposing an individual's head onto someone else's body, is similarly getting easier, although it still requires an actor or a body double to perform body actions and impersonate mannerisms.[108] Similar advances are taking place in synthetic audio.[109]

Some researchers are skeptical of the deepfake hype and suggest that their weaponization will depend on pragmatic considerations, such as ease of access, decreased cost, and diminishing effectiveness of current techniques that are not ML-enabled. They argue that as long as less automated techniques produce content that can go viral, operators have less incentive to invest in deepfakes.[110] Well-resourced operators could plausibly integrate deepfakes into their arsenal, but decisions to deploy may depend on risk tolerance—the risk of being discovered weighed against an unclear reward. Defensive forensic capabilities increasingly focus on quickly detecting deepfakes of national leaders and high-profile public figures, which can further deter potential deployment. Nevertheless, deepfakes may be incorporated in disinformation operations in several plausible scenarios.

Deepfake video and audio may become more prevalent on newer platforms, where engaging content can go viral quickly and there is limited verification that speakers are who they claim to be. For example, a person claiming to be Brad Pitt drew thousands into a

conversation on Clubhouse, before he was identified when someone became suspicious of his voice.[111] Researchers at Mandiant already demonstrated how fine-tuning a publicly released model from a few voice samples can produce a cell-phone quality audio of Tom Hanks saying what researchers prescribed him to say.[112] Malicious operators using samples of less recognizable voices than world-renowned actors, such as a military leader, a member of Congress, or another government official, could plausibly create audio impersonations that amplify disinformation. Disinformation operators could target newer platforms that are optimizing growth over security to test new techniques and fine-tune their efforts.

Threat actors may also integrate tailored deepfakes into a hack-forge-leak type of disinformation campaign. A "leaked" deepfake video of U.S. or allied troops committing atrocities, for instance, has long been discussed as a possible catalyst for unrest or international crisis. Disinformation operators could deploy deepfakes against diplomats to sow discord among allies.[113] In a plausible election disinformation operation, a tailored deepfake of a political candidate or their family member delivering extreme comments could surface online as part of an alleged hack and is then "leaked" in the weeks leading up to an election. Funneling this material through domestic political actors would further allow threat actors to exploit the heightened partisanship of the election cycle and the limitations in the ability of U.S. authorities to investigate constitutionally protected speech by domestic actors. Such a disinformation campaign could potentially sway undecided voters in key swing states. Even debunked, this type of content is engaging enough to go viral, exposing millions to disinformation before it is detected and throttled by social media platforms.

Finally, and most impactfully, the slow drip and normalization of deepfakes may lead to an overall erosion of trust. In recent years, the number of deepfake videos has grown exponentially, doubling every six months.[114] While a disinformation campaign using deepfakes would likely be discovered eventually, their growing popularity may erode trust in video evidence of wrongdoing. This "liar's dividend" effect—the dismissal of a truthful occurrence as a deepfake—may further erode societal trust in government

institutions and democratic process.[115] In a polarized society, the proliferation of deepfakes may pose a pernicious challenge. The long-term erosion of trust and inability to separate fact from fiction at a societal level can exacerbate the cycle of cynicism, creating fertile ground and increased vulnerability to disinformation, with or without deepfakes.

Enhance and Deploy Dynamic Bots

As the social media platforms improve their detection capabilities, the life spans of inauthentic digital armies and their windows to blast disinformation are getting smaller. But ML can help the next generation of social bots become better at mimicking human behavior and achieve maximum amplification in shorter time frames, augmented with ML-generated avatars.[116] For example, researchers proposed to enhance social bots by extracting features from text and image content and processing them through neural networks to allow a social bot to publish contextually relevant comments.[117] While research literature theorizes the evolution of social bots with ML-enabled capabilities, open-source repositories and dark markets so far offer only limited end-to-end solutions.[118]

Meanwhile, AI developers were among the early innovators to use social media platforms as a testing ground for sophisticated bots.[119] The techniques of bot detection services can be exploited to help build bots that behave more like humans. Supervised learning techniques use data curated by humans to teach the machine what is a "human" and what is a "bot." Threat actors can reuse the training datasets released by bot detection services, which examine features of a profile, number of followers, time of activity, language, and sentiment to classify a Twitter account as "likely bot" and "likely human."[120]

In addition, unsupervised ML can help build more human-like bots. In theory, a model trained through unsupervised learning techniques to draw insights from data could distinguish between human and automated patterns in how accounts connect with each other and spread information. This can help threat actors develop

social bots that better imitate the human pattern.[121] In an attempt to improve bot detection systems and test their limits, researchers are exploring these approaches to build bots that would beat the bot detection systems and get classified as human.[122] They aim to point out weaknesses in existing bot detection systems and improve them. Yet this line of research could also show threat actors how to build bots that are indistinguishable from human-operated accounts online.

Some sophisticated bots can already emulate human patterns of posting and consuming content.[123] It is an open question, however, whether this will achieve the same level of virality as current computational propaganda techniques. Bot operators may lose reach as they try to emulate humans, unless they create content that is engaging enough to go viral organically.

ML techniques may soon enable social bots to maximize amplification and evade detection with varied synthetic content. Simpler versions of social bots tended to post pre-scripted set of phrases and keywords programmed into their code with a pattern of repetition that can get flagged by platforms. If synthetic text models can increase the variety of terms and phrases to resemble normal human speech, the bots may fool automated defensive systems.

Social bots may not replace humans altogether. Rather, an advanced social bot system could combine already-available building blocks into a human-machine team or an entirely automated solution. Human operators can theoretically team a social bot with social listening tools and ML generative language systems.[124] First, this system could detect trending topics and hashtags through social listening tools and decipher the sentiment with a sentiment analyzer. Then, human operators could instruct the language generation component with prompts and examples to generate posts optimized around the trending subjects, a target audience's identity, and the disinformation campaign's narratives. This system could generate a pool of varied messages that appear authentic and relevant to the real-time online conversation. Next, a human could curate the most relevant outputs and feed these into

a pool of messages. Finally, the social bot can automatically pull messages from the pool and post them.

Threat actors could plausibly develop such a system and, with some engineering ingenuity, automate human curation out of the loop entirely. Given the current means by which most ML language generation systems work—through prompts—fully automating an advanced social bot that will dynamically generate its own tweets and hijack trending topics in real time may be difficult. However, it will become more probable as more developers make their language models freely accessible. If threat actors do not care about the quality of the output and are willing to risk the social bot occasionally posting nonsensical statements, they could risk automating the social bot end-to-end.

Finally, more plausibly, the next steps in social bot evolution to mimic human behavior may come from the capability to respond dynamically and interactively as social bots integrate with chatbots, an AI-enabled technology we examine in a later section.[125] The above scenarios assume that the same type of API access that the platforms expose to "the good bots" for legitimate uses is also available to bots that are able to mask their true nature from the platform's vetting process.

Boost Social Engineering of the Super-Spreaders

ML systems may enhance social engineering techniques for targeting influencers, or "super-spreaders," who can unwittingly amplify the campaign's messages. Threat actors have already successfully impersonated experts using inauthentic personas and duped news outlets with large followings.[126] By assembling a dataset of posts scraped from accounts of public figures or acquired from the platforms through API access, threat actors could apply sentiment analysis and stance detection techniques to identify frequently discussed topics that super-spreaders reshare. They could then generate synthetic content similar to collected posts to entice the super-spreader into unwittingly resharing the new content.

While some of this is speculative, all the building blocks are available today. The amount of engineering effort may seem prohibitive to a small disinformation outfit or overkill to get a specific influencer to retweet a meme. But for an advanced state-sponsored threat actor, this is a long-term investment. Multiple high-profile figures can be cultivated with ML-enabled tools, assuming they manage their social media presence themselves. A series of posts from a super-spreader can reach millions of followers, receive additional amplification from broadcast media, and spread the threat actor's messaging organically. The effort pays for itself when the disinformation narrative becomes endemic.

Exploit the Recommendation Algorithms

Disinformation actors can amplify their campaign content to new audiences by taking advantage of a platform's recommendation algorithms. Powered by ML, the recommendation algorithms infer a user's interests and viewpoints to deliver tailored content. Every user interaction with a "heart" or an "angry" reaction can teach the algorithm about user preferences, a process known as "self-selected" personalization.[127] The algorithm then pushes the content to another user who is either connected or is a "look-alike," fitting the profile of someone who might also like the content—a process referred by some researchers as "pre-selected personalization." Both self-selected and pre-selected personalization can create a so-called filter bubble, a closed information loop which can increase polarization and drive users towards fringe content.[128]

The volume and simultaneity of data posted to social media platforms requires them to decide what content to prioritize showing to the user, a ranking process that ML helps to manage. At the core of many social networks is a system based on deep learning recommendation models (DLRM) and the concept of collaborative filtering—the idea that the best recommendations come from people who have similar tastes. Collaborative filtering prioritizes the content—posts, pages, groups, events—that a user might like based on the user's reactions and those of people with similar profiles. Facebook's average dataset for collaborative

filtering has 100 billion ratings, over 2 billion users, and millions of items—a massive scale.[129] Its Newsfeed algorithm ranks the relevance of content to the user based on multiple criteria, including personalized data from past activity, information about the post's author, and the popularity of a post. Other platforms also collect massive amounts of user data to tailor content. Platforms can tweak recommendations by altering the relative importance of various criteria.[130] Most platforms that recommend content use similar deep learning architectures.[131]

These types of architectures can exacerbate filter bubbles that threat actors can exploit. As one former YouTube engineer explained, as an algorithm gets better at predicting who may find content engaging, it is less likely to recommend such content to those who will not.[132] Algorithmic recommendations, such as promotion of borderline and manipulative content, become harder to notice, because the content becomes less likely to reach users who might be offended and report it to the platforms.[133] In some instances, recommendation engines can gradually increase the polarization of content they suggest, increasing controversy enough to keep users engaged, but not enough to be reported as violating terms of service.[134] Researchers at DeepMind similarly concluded that feedback loops in recommendation systems can feed filter bubbles and create echo chambers, narrowing users' exposure to diverse content and potentially "shaping their worldview."[135]

Threat actors can also hack the recommender algorithms. On the less sophisticated end of the spectrum, "shilling attacks" against recommender systems can simply involve creating fake user profiles that friend authentic users or join authentic groups, amplify content with a high volume of likes and engagements, or swarm the posts supporting the narrative with positive ratings while attacking opposing views.[136] These actions, in turn, increase the exposure of the target audience and their algorithmically profiled lookalikes to malicious content. These techniques may not use ML, but rather abuse the ML within the recommender systems.

The second type of attack on the recommender systems targets "data voids," or gaps in content returned by the search engine and

social media search functions in response to specific search terms.[137] Examples of data voids include strategic new terms or outdated terms. Once threat actors identify obscure search queries that return sparse content, they can create content that populates when these search terms are entered. ML techniques can help threat actors find these voids by varying search terms and filling them using a natural language generation system to create malicious content. Other amplification techniques can drive users to search for specific terms, lending credibility to deceptive messages. Search-adjacent recommendation systems such as auto-play, auto-fill, and trending topics are similarly vulnerable to manipulation.[138]

Finally, threat actors can harness adversarial approaches to amplify their content. Like other ML systems, recommendation algorithms are vulnerable to evasion attacks.[139] In this type of attack, an operator makes subtle changes to an input that, while imperceptible to a human eye, causes an ML system to make a mistake. In a proof of concept, security researchers have applied this technique to recommendation systems that decide what content a user sees on their social media feed. Researchers manipulated the recommendation algorithms by gradually modifying genuine images to fool a classifier and cause it to make an incorrect prediction.[140] In this attack, introducing small imperceptible changes to images in a post can significantly increase a post's relevance score and improve the rankings of related items by the social media recommendation algorithms. The higher ranking pushes the post with the manipulated image to more users' feeds, amplifying its reach.[141]

In theory, similar attacks could exploit both the content and the group recommendation algorithm to make it easier for users to discover groups controlled by the disinformation operators. A group's cover image can be manipulated with the attack described above. For example, when Facebook adjusted its recommendation algorithms to weigh community content more heavily, it prioritized engagement with most relevant content and social interactions with friends and family through groups.[142] In addition to other unintended consequences, such as prioritizing polarizing content, this change also increased group discoverability, making it easier

for users to find fringe groups.[143] However, in response to the rapid growth of conspiracy theory communities such as QAnon, the platform recently removed civic, health-related, and overtly hyper-partisan groups from its recommendation algorithm. Yet this change did not impact groups in other categories. [144] Using the attack described above, threat actors can manipulate groups' cover images to increase the ranking of their group by the recommendation algorithm and drive traffic to the seemingly innocuous groups they cultivate, such as private interest-based groups about tourism or food. For example, once they have built their audience on beautiful pictures of Riga, as we describe in our RICHDATA Framework report, they can repurpose the group for a future disinformation campaign.[145] This technique can potentially turn innocuous private groups into self-reinforcing echo-chambers for radicalization, hard for researchers to observe and for platforms to police.[146]

Barring further interventions, recommendation algorithms trained on the preferences of like-minded individuals and tuned to reinforce users' leanings are likely to perpetuate polarization that threat actors can deepen and exploit. Despite recent experiments to reduce exposure to extreme and malicious content,[147] platforms continue to state that users remain in charge of their experience.[148] This means that user biases—and choices by their algorithmically profiled look-alikes—will continue to drive filter bubbles and echo chambers.

Boost Conspiracy Information Laundering

In 2018, Mark Zuckerberg noted that Facebook research "suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average."[149] Users find borderline content more engaging, and recommendation algorithms are built to serve users the content they find most engaging, underscoring and exacerbating the participatory nature of disinformation.

ML can help threat actors exploit this feature to wrap their message into engaging conspiracy theory content. Researchers studying the growth of anti-vaccination and the QAnon conspiracy communities warned that threat actors affiliated with Russia cultivated, amplified, and aligned narratives within these communities.[150] In theory, trained ML models could find patterns across conspiracy narratives, to replicate and scale the type of cross-pollination that occurred between QAnon and the anti-vaccination community during the COVID-19 pandemic.[151]

Finally, threat actors can use natural language generation systems to produce new conspiracies and seed new narratives. Researchers have demonstrated the ability to generate synthetic text mimicking the style and content of QAnon messages.[152] Threat actors could fine-tune an open-source large language model on a dataset that contains different conspiracy theories and recent news stories, to then generate a new set of conspiracy messages that appeal to broader audiences or even replace existing players.[153]

Automate Trolling

Content generation capabilities within conversational AI technologies may soon increase the scale of trolling. Conversational AI is an umbrella term for technologies that enable machines to communicate with individual users in conversational language through voice, text, and video. They use large volumes of data, ML, and NLP to imitate human interactions, recognizing speech and text input and generating a response.[154]

Open domain chatbots may soon reduce the need for human trolls, while natural language generation can improve the output of automated trolling. These systems are conversational AI agents that can engage in continuous dialog on a variety of subjects in multiple languages and accomplish decision-making tasks.[155] In 2020, the Brain Team from Google Research presented Meena, a multi-turn open domain chatbot trained on data from public domain social media conversations.[156] Meena could engage in convincing conversations—she made jokes, spoke colloquial

English, and returned relevant answers to questions across a variety of topics.[157] Compared to her predecessors, Meena provides a glimpse into the future of chatbots and their growing sophistication. [158]

Conversational AI platforms are also becoming more accessible as the advanced building blocks for creating chatbots enter the open-source market. [159] Many high-performing chatbots developed by well-resourced organizations for legitimate purposes are open source, such as Facebook's BlenderBot, the first chatbot model to exhibit humanlike traits, assume a personality, maintain a lengthy conversation, and show empathy.[160] Researchers released the complete model, code, and evaluation set-up to drive research forward.

These capabilities are also diffusing as countries race to claim supremacy in AI. Chinese technology giant Baidu recently built its own version of a chatbot. PLATO-XL, a 10 billion parameter conversational model, can engage in multi-turn dialog in Chinese and English. This diffusion is likely to continue as technology matures and costs reduce.[161]

For threat actors these capabilities may be hard to resist, particularly as part of a human-machine team. A chatbot drawing on advanced large language models could be a believable character on a variety of platforms with reduced need for a human operator. Using fine-tuning, threat actors could train a chatbot to specialize in specific trolling techniques, such as provocation, nuisance, or social engineering.[162] With ML tools able to "understand" online conversations and respond in natural ways, chatbots could present a variety of responses from which a human operator could select, curate output, and post a response. This would allow a human operator, especially one operating in a foreign language, to manage conversation streams more efficiently and effectively. Theoretically, a human operator could run multiple "plays in three acts" simultaneously, with multiple chatbots role-playing an argument among themselves to manufacture consensus or stir up controversy.[163] Whatever their objective—to support a viewpoint, to attack users expressing alternative views, or to generate engagement—chatbot trolls could become increasingly

difficult to distinguish from humans.[164] Human-machine teaming can reduce the labor-intensive process for the solitary human operator, help more closely impersonate authentic users, and engage more humans in online conversations through an exponential increase in scale.

Rapid advancements in conversational AI could plausibly lead to fully automated trolling, removing the human operator from the loop. An automated trolling system could leverage several technologies: chatbots engaging in multi-turn conversations in comments on posts, natural language generation systems to write cogent paragraphs advancing an argument, and ML-powered sentiment analysis to assess the effect of their messaging and optimize deployment to comment sections of specific posts. Built into the trolling chatbot, controversy detection—another subfield of sentiment analysis—could pick up on underlying controversy not explicitly mentioned in text and help identify new wedge issues to exploit.[165]

That said, platforms have a role in limiting some of the worst effects of these scenarios. Autonomous deployment of a trolling chatbot system on social media would require access to the platform APIs. In addition, a trolling system would have to masquerade as a legitimate application, such as an airline agent chatbot, as platforms are unlikely to allow an illegitimate trolling chatbot access to the platform. Threat actors' success in this effort substantially depends on what chatbot behavior the platforms deem legitimate, and the ability of operatives to evade the platforms' detection methods. Mitigations may be more difficult if conversational AI systems are built into the social media platforms themselves and then exploited.

Scale Up Mobilization with Personalized Disinformation

NLP, open domain chatbots, and deepfake technology may advance the process of mobilizing a targeted population. ML applied to the actualization stage allows the creation of more precisely targeted and personalized disinformation to cultivate

influencers and inspire them to action. Threat actors can combine the various techniques we have discussed previously to enhance the actualization of their campaign.

The built-in micro-targeting features of platforms, social network analysis, and sentiment analysis of an individual's posts, could feed precisely targeted disinformation and focus it on those users that are most likely to take action. ML-enabled social network analysis can identify similarities between individuals and groups and gauge the strength of their connections. Advanced sentiment analysis and psychographics systems can help build a detailed personality profile from language.[166] A stance detection model could identify users exhibiting signs of disillusionment, cynicism, susceptibility to influence, and aggression, helping operators target them for cultivation.[167] One study found that violent language had an impact on "aggressive citizens," where even mild violent metaphors significantly multiplied support for political violence in individuals displaying this trait.[168] Natural language generation systems can help human operators hone the unique lexicon of fringe communities and equip the operator with a stream of convincing radicalizing content to drive individuals to real-world action.[169]

Taking this a step further, advances in chatbots and deepfake technology may soon combine to give threat actors new techniques to engage individuals convincingly and persuade them to action. Efforts like the Russian IRA's operations to organize rallies in the United States could be supercharged through personalized chatbots and deepfake impersonations as operators upgrade direct messaging to engage unwitting targets over video chat.

A personalized chatbot that "knows you" may soon be a tool in the disinformation operators' arsenal. Chatbots are increasingly more expressive and can respond naturally using everyday language.[170] By building what feels to humans like genuine relationships, chatbots are becoming companions and trusted confidantes.[171] In theory, threat actors could fine-tune or draw on open-source models to build personalized chatbots that serve their cultivation purposes. Creating and training simpler chatbots is already publicly available and requires no knowledge of ML.[172] Chatbots fine-tuned

on the dataset of personal messages and posts of a loved one are already interacting with humans via text and audio.[173] Operators could apply the same social engineering techniques used against super-spreaders to build custom datasets for lesser known targets of interest, such as grassroots activists, someone engulfed in a conspiracy theory community, or leaders of an extremist organization. Scraped public posts from targeted individuals and their close network could fine-tune a chatbot on the language pattern of an acquaintance. Set in a context of presumed trust, a personalized chatbot that sounds like a distant colleague can connect with the target through a social media direct messaging to organize events on the ground, while obscuring any red flags of the foreign disinformation operator's origin.

Finally, threat actors may soon have powerful options to scale up cultivation and mobilization by bringing chatbots to life with synthetic video skins for video calling. In 2020, a researcher at the BlackHat cybersecurity conference, using an open source chatbot and deepfake technologies, created a synthetic clone of himself. He fine-tuned a chatbot on his own text messages, combined it with a synthetic video skin of himself, and had his "clone" carry on a video-call with a friend.[174] While a glitchy prototype at the time, a similar technique could help threat actors increase the scale of their operations to engage many susceptible users simultaneously and move them towards action. Technology is evolving to make impersonation possible in real time over a video call, enabling threat actors to mask their origin as they cultivate their targets with direct engagement.[175] Avatarify is just one example of an open-source application that provides a digital avatar of someone else's face over one's own in video conferencing applications.[176] This solution is another implementation of a head-swapping deepfake technique discussed in the earlier sections and only a few images of the target are necessary to impersonate them—a sample easy to acquire from their public social media profiles. The technique automates image animation by combining the appearance extracted from the target's image to impersonate them with motion patterns derived from the real-time video of the impersonator.[177] At present, Avatarify requires additional code to run and hardware with enough computing power for video-gaming. These are not

obstacles to a skilled and well-resourced operation. In addition, technology is likely to evolve to run on more accessible hardware. Network lag and low lighting can help mask any observable lip-syncing issues. The source-code for this app is readily available on GitHub for threat actors to experiment with and improve.[178]

What sounds like science fiction—a chatbot that talks like a trusted colleague, an imposter donning a synthetically-generated face of a friend in a real-time video call, or some combination of both—may soon emerge as the next set of disinformation tools. Threat actors leveraging advances in ML that enable chatbots with synthetically generated skins could become more successful in building online relationships and persuading humans into action or inaction offline. Chatbots today may not yet be nuanced enough to effectively radicalize a human into creating real world effects, but this may not stay true for long.

Key Findings

> *"If everybody always lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer. And a people that no longer can believe anything cannot make up its mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please."* [179]
>
> *– Hannah Arendt*

The application of ML to disinformation operations is rapidly evolving, powered by data abundance, innovative algorithms, and massive computing power. Authoritarian regimes, notably China and Russia, are actively pursuing AI/ML research in a number of areas that will enable greater manipulation of their domestic information environments, while directing their disinformation capabilities at adversaries outside their borders.[180] Open societies are vulnerable to the malicious effects of AI-powered influence as they struggle to balance free speech with the harmful effects of disinformation designed to provoke, disrupt, and divide. Resolving this tension in the pluralistic marketplace of ideas may be increasingly difficult in the atmosphere of a fractured information environment and personalized information loops.

ML technologies are likely to exacerbate the current disinformation crisis. Disinformation campaigns exploit the features of today's information environment and the deeply embedded characteristics of human cognition. Presently, social media users produce and consume an immense amount of content. In 2020, humanity posted an average of 500 million tweets, viewed 4 billion Facebook videos, and created 4 billion snaps—every day.[181] Human cognition is ill-equipped to handle the volume of content that the average user is exposed to daily. The potential deluge of AI-powered disinformation may strain individual decision-making, societal cohesion, and the ability to reach consensus that are

critical for democracies to function. Policy and technical countermeasures are necessary to mitigate the disruptive effects, in a whole-of-society effort.

Protecting the shared public square requires an understanding of the threat of AI-powered campaigns. The below findings outline how AI/ML capabilities may exacerbate existing trends and aid disinformation campaigns, followed by recommendations to mitigate their impact.

AI will augment stages of disinformation campaigns, but the gains are uneven

ML can assist human operators in creating more engaging content and designing more precisely targeted disinformation campaigns. In the interactive stages of troll patrol and actualization, ML can currently augment the work of human operators, but full automation at scale would require significant software engineering effort, further ML advances, and nearly unfettered access to platform APIs and user data. The degree of success depends on how well operators mask their true intentions.

Furthermore, ML can also assist in the reconnaissance stage by providing a more nuanced understanding of the information environment through social listening and sentiment analysis. It can help identify societal fissures and fine-tune personalized disinformation from the data that individuals and target societies post online. AI-generated fake personas will assist in the task of building infrastructure and make the detection of fake accounts and campaigns more difficult. Newer ML technologies, particularly natural language generation and visual synthetic media, are now capable enough to help human operators create engaging and targeted content that can go viral.

Social bots are likely to pass as humans online, while conversational AI chatbots will enhance various trolling techniques. Donning AI-generated clones of trusted persons and armed with insights from user data, advanced multi-dialog chatbots may further personalize disinformation, engaging humans one-on-one and mobilizing them to action. From artificially generated avatars,

resumes, and employers to fake pets, food, and apartments, nonexistent humans are poised to live believable lives online. However, full automation at scale will require software engineering effort and further AI advances before it becomes a significant threat. Given the benefits, well-resourced actors are likely to invest in AI-enabled automation to augment human operators and increase the speed and scale of their operations.

Foreign, domestic, and transnational actors may drive ML-enabled disinformation

"Disinformation starts at home" is a popular axiom among disinformation researchers. Nation-states often test new disinformation tactics on their domestic audiences or countries within their perceived sphere of influence before applying them more broadly. As more nations apply ML-enabled capabilities, they may test them domestically first—a trend to watch. When disinformation campaigns move abroad, they often exploit grievances and fault lines within the targeted country. By amplifying existing content, such as conspiracy theories, they can mask their activity and further undermine societal trust.[182] The lines are also blurring between foreign and domestic disinformation, as operators employ "influence as a service" firms to mask their activity, making attributions and countermeasures more difficult.

While foreign disinformation campaigns continue to command attention, domestic actors are adopting similar tactics in political campaigns. More than 65 firms in 48 countries deployed computational propaganda services on behalf of political actors spending almost $60 million since 2009.[183] Of the 150 operations in 50 countries that Facebook disrupted in the past 4 years, nearly half targeted domestic audiences.[184] The 2020 U.S. presidential election saw domestic political actors using inauthentic accounts and the volume of domestic disinformation dwarfing foreign influence efforts.[185]

AI-driven advances have attractive applications for political campaigns. Despite the lack of measurements to prove its efficacy, the case of Cambridge Analytica demonstrated a proof of concept in how ML-powered micro-targeting and psychographic profiling

could draw on private and public data for influence. Today, social media companies often remove election-related activity on the basis of actor behavior, such as inauthenticity, coordination and repetition of posts, rather than the content itself.[186] De-platforming on the basis of what users post is reserved for specific platform policy violations, such as content inciting violence.[187] While platforms are getting better at identifying inauthentic accounts, domestic actors—or the "influence as a service" transnational firms they hire—could use ML-powered language generation systems to spread disinformation through authentic users and influencers, straining content moderation efforts.[188] These systems enable operators to generate a greater variety of messages to advance a talking point, avoiding copying and pasting techniques, known as "copy pasta," that can be easily flagged by platforms as a sign of coordinated activity. In 2019, Twitter's then-CEO sounded the alarm about the impact of AI on political discourse, warning: "Internet political ads present entirely new challenges to civic discourse: machine learning-based optimization of messaging and micro-targeting, unchecked misleading information, and deep fakes. All at increasing velocity, sophistication, and overwhelming scale."[189] As AI-enabled content generation tools become more widespread, the temptation to use them by candidates or elected officials will only grow, leaving social media platforms in an unenviable position of having to make nuanced moderation decisions about a high volume of protected political speech and risking accusations of bias.

While political speech is protected in the United States, questions remain about whether these protections should extend to AI-generated or bot-amplified speech. With a broad interpretation of protected speech, AI-powered disinformation at scale has the potential to reach large audiences and become endemic. Developing norms to discourage the use of artificial amplification by political campaigns and their affiliates during elections is not a panacea, but a necessary start.

Social media platforms and ML applications serve as both the battlegrounds and the weapons of disinformation campaigns

Social media platform features define the terrain and shape threat actors' tactics. Disinformation operators will seek to misuse social media platforms and their features, exploiting their design that helps people to connect and engage. As long as market incentives drive growth over security, there is a risk that the architecture of established and new platforms will be initially insecure and difficult to harden later.

Adding to the challenge, many ML applications are dual-use. One person's humorous synthetic video is another's deepfake. These technologies are driving innovation in a variety of industries, from increasing the supply of stock photos for digital marketing to bringing life-like avatars into virtual environments for war-fighter simulations and remote workspaces. However, deepfakes are now used in head-swapping apps for nonconsensual pornography, as a tool of political warfare, and to subvert detection of fake accounts by platforms. NLP in chatbots augments service centers and increases the efficiency of virtual assistants. However, it also opens the door to increasingly advanced chatbots that carry out precision trolling. ML systems using advanced psychographic targeting help advertisers expand their audiences and enable political campaigns to identify undecided voters. Malicious actors can use these same tools to micro-target and mobilize unwitting audiences in a disinformation campaign.

The open nature of AI research provides avenues for misuse

The open-source culture of the AI research community has fostered many recent advances, but it also provides ready access for malicious actors. Top AI research institutions and independent researchers post source code to GitHub to share with others.[190] There are tutorials about how to fine-tune the open-source code to generate synthetic text with a specific tone, or how to create a lip-synching video of yourself speaking 30 different languages.[191] The many benefits of open-sourcing include generating public awareness, enabling researchers to build on shared research to drive innovation, and generating training data to improve deepfake

detection systems. However, efforts to increase public interest in technology, such as a website with images of convincing GAN-generated fake people, provide less sophisticated threat actors ready access to this technology.[192] Threat actors can fine-tune open-source ML models to produce hyperrealistic images or clone voices of specific individuals with precisely curated small training datasets with relatively low computing requirements.[193]

This tension is reminiscent of the offense-defense debate in the cybersecurity research community. The default assumption of those who advocate for openly publishing nearly ready-to-use ML models is akin to security researchers releasing information about newly discovered vulnerabilities and publishing exploit code to force software vendors to fix them.[194] Both are motivated by the goal of improving security and advancing research, yet they can invite threat actors to adopt the tools and exploit the vulnerabilities, leaving unwitting users vulnerable until defenses catch up.

Disinformation campaigns cross platforms, while technical and policy countermeasures are platform-specific

Disinformation campaigns in their present form are asymmetric. Threat actors often operate across many platforms making it more difficult to understand the full scope of activity. Companies work independently to counter malign activity on their platforms; however, building visibility of the larger campaign requires greater coordination among targeted entities. Despite some automated threat detection on larger platforms, threat hunting remains a manual and labor-intensive process that relies on tips from the government, the media, and civil society partners. Silos between platforms, a lack of standardization on sharable information, and the varying degrees of threat monitoring and self-policing between large and small platforms impede discovery, neutralization, and attribution of cross-platform campaigns in real time.

In addition, the lines are blurring between the foreign-manufactured disinformation and the homegrown misinformation ecosystems. Threat actors increasingly seed their payloads on small fringe platforms or entirely "off platform" on their own websites, where it is below the radar of the larger platforms' threat

hunting teams. From there, the disinformation finds its way into the mainstream platforms, grows, and retreats to the fringe to regroup after takedowns by major platforms. While large platforms deploy AI-enabled tools and human threat investigators to detect behavior that violates their policies, smaller platforms often lack the requisite tools, human capacity, or the will to police content for harmful use.

AI-enabled recommendation algorithms can fuel personalized disinformation

By tailoring each user's information diet based on their past engagements and the actions of algorithmically determined similar profiles, recommender algorithms can move their users into closed-loop information systems further exacerbating polarization.[195] Over time recommendations lose their effectiveness and require continuous intervention to prevent the narrowing of the pool of content presented to the user. Recommendation algorithms can also be manipulated by disinformation actors to exacerbate information silos, driving users deeper into extreme groups and conspiracy theories.

While social media platforms battle this trend with various technical solutions, it remains difficult to independently verify if their methods are working.[196] The scale of today's information environment means that platforms must make choices about what data to surface to their users. At the crux of the problem is the tension between putting users in charge of their experience and adjusting the algorithmic architecture that exacerbates those choices.

Traditional media are targets of AI-powered disinformation campaigns

Media coverage of disinformation, misinformation, and hacked-and-leaked information plays a critical role in its virality and reach. A notable example is the media amplification of hacked and leaked Democratic National Committee and Clinton campaign emails.[197] Because of the reach, media organizations are prime targets of disinformation campaigns. However, they often lack the necessary tools to check the authenticity of ML-generated information and its

sources. Synthetic media can further obscure authorship, making the task more daunting. While government organizations and technology companies are developing tools to detect synthetic image, video, and audio, few capabilities exist for the detection of machine generated text. Some media organizations have begun to put in place safeguards to avoid becoming super-spreaders of disinformation, but this practice is not universal.

Digital media literacy is critical in countering ML-powered disinformation

While companies attempt to detect disinformation operations on their platforms, a substantial amount of malicious content still reaches its intended audience. Most solutions focus on mitigating the supply of disinformation. Relatively less effort has focused on the demand-side—the cognitive biases, individual grievances, and digital media literacy of individual users. The factors that increase societal vulnerability to disinformation require urgent attention. Chief among them is the need to build societal resilience and awareness about ML-generated content and the techniques disinformation operators use to exploit users' data in order to shape their perception of reality.

Recommendations

Jake Sullivan, President Biden's national security advisor, recently warned of the new challenges of disinformation and cyber intrusions from the technologies once thought would almost inevitably favor democratic values. Sullivan said, "if democracies don't turn back this tide, the second phase of the digital revolution will grow darker with the proliferation of autonomous disinformation."[198] The challenges engendered by these campaigns were already present without the addition of ML techniques which is why the United States must prepare for the escalation of ML-powered disinformation campaigns. Like other complex issues there are no simple solutions. The recommendations below target key stakeholders that play a role in moving us towards a healthier information ecosystem in the future.

Develop technical mitigations to inhibit and detect ML-powered disinformation campaigns

The AI-enabled disinformation challenge cannot be solved solely by technical approaches, but technology can do more, particularly in combination with policy mitigations. Select mitigations include limiting threat actors' access to user data and detecting and labeling AI-generated content and systems.

Inhibit access to user data by threat actors and their proxies

Disinformation operators rely on user data to build their campaigns and train ML systems. Platforms should restrict access to user data through their APIs. Congress should regulate what user data is authorized for sale through data brokers. Specifically, Congress should limit services that aggregate consumer and personal information from public records, link them to specific users, and sell this access to clients without vetting the purchaser. It should set consistent rules across platforms for sharing data and requirements for vetting third party applications.[199]

Develop interoperable standards for detection, forensics, and digital provenance of synthetic media and labeling of chatbots

Technology for detection and forensic analysis of AI-generated content is locked in a race with the technology that creates them; as detectors improve, so do the generators, leapfrogging one another. While promising industry and government efforts at detecting synthetic audio-visual media are emerging, few solutions are underway to detect synthetic text.[200] Detecting ML-generated content would enable timely mitigations including labeling and removal as necessary. Identifying and labeling chatbot systems would allow humans to know when they are engaging in a conversation with an AI-enabled system. Standardized methods to detect synthetic media will allow greater transparency across the platforms. Government funding can boost studies into the development of novel techniques for synthetic text detection and should support current industry initiatives, such as the Coalition for Content Provenance and Authenticity, to jointly develop technical standards of authenticity and provenance for synthetic content that can be built into the online environment.[201]

Develop an early warning system for disinformation campaigns

Disinformation campaigns exploit human and digital networks and require a networked defense. Disinformation operators have tended to hone their craft in culturally familiar information environments against domestic opposition and within nations in their perceived sphere of influence. The U.S. government counterpropaganda and disinformation community should monitor these testing grounds closely to detect early indicators of new campaigns and techniques, including experimentation with AI technologies. Federal government efforts should leverage allies and partners to share information and best practices on detecting disinformation campaigns.[202] The U.S. government should consider expanding intelligence sharing on disinformation operations to non-allied partners on the frontlines of adversary experimentation. U.S. missions overseas also need the tools and the training to identify new disinformation techniques surfacing in regional campaigns that may impact the U.S. domestic information

environment. A process to share the signals of developing campaigns can assist in the early detection of these threats and empower rapid government responses.[203] For disinformation campaigns in the United States, cross-platform collaboration and information sharing is key in creating a shared common operational picture. Models such as the Information Sharing and Analysis Organizations can facilitate communication between government, industry, and researchers to increase situational awareness.[204]

Build a networked collective defense across platforms

Disinformation campaigns increasingly cross platforms, while technical and content policy countermeasures are platform-specific and siloed. There is a need for greater understanding of how campaigns develop and move across platforms. Information-sharing between platforms on emerging or ongoing disinformation campaigns is largely informal. Compounding the challenge, there are varying degrees of transparency and sharing of disinformation content with academic and civic society researchers by the platforms. Coordination on content moderation, exposure, and removal of disinformation campaigns is ad hoc, despite the perception that platforms sometimes follow one another's lead in potentially controversial cases.[205]

Cross-platform capabilities currently advantage digital marketers and disinformation operators. Tools to simultaneously propagate influence messaging to multiple social media platforms are readily available. However, there are few tools that do the opposite, such as those that identify disinformation campaigns across different platforms. A patchwork of experimental solutions may be emerging in the national security community, but they focus on increasing awareness in the foreign information environments and cannot be used in a U.S. domestic context.[206] A nongovernmental effort should address the gap where foreign disinformation operations grow roots in the target society, mix foreign disinformation and domestic misinformation, and operate across platforms. Social media platforms can fill this gap by building partnerships to prepare for the emergence of ML-enabled disinformation campaigns. Technology platforms and civil society should invest in solutions to increase situational awareness in the domestic

information environments to understand the scope and scale of ML-powered campaigns.

Mitigations should focus on increasing transparency and accountability, removing impediments to sharing threat information, and formalizing mechanisms for collaboration. Social media platforms regardless of size should have policies, processes, and staffing to disrupt disinformation operations and publish regular reports about disinformation operations on their platforms. In addition, the lack of comprehensive privacy regulation and standardized rules for vetting and access to data impedes outside researchers who track disinformation campaigns across platforms. Platforms often limit the information they share due to privacy concerns and inconsistent vetting processes. Congress should remove impediments to sharing threat information and fund efforts to establish standards for the sharing of technical metadata with vetted researchers, the government, and across platforms.[207] Finally, to neutralize cross-platform disinformation campaigns, the platforms and researchers should establish a formal collaboration mechanism to enable early warning of developing campaigns, to share best practices, and to empower rapid response. Examples of cross-industry efforts exist in other problem domains, such as cybersecurity and counterterrorism, and can provide models for sustained collaboration on an ongoing basis.[208]

Examine and deter the use of services that enable disinformation campaigns

Private companies offering marketing, PR, and consulting services—"influence-as-a-service" entities—are increasingly a tool of disinformation operators that use automated amplification techniques and deceptive practices to influence online discourse. The market for these services is growing, ranging from harvesting of user data to selling inauthentic accounts, likes, and comments to inflate of the size of their clients' followings. These services are likely to leverage ML capabilities such as natural language processing and content generation as the underlying technologies become more accessible. Users' digital footprints on and off social

media platforms can combine to provide a powerful dataset for ML-powered insight generation and targeting.

Congress should examine the current use of ML-enabled tools and deceptive techniques by firms providing influence for hire and build norms to discourage their use. Developing norms around these practices is the first step toward raising awareness and deterring it, before ML-enabled content generation tools further increase the scale, dilute the authenticity of public discourse, and undermine trust.

Integrate threat modeling and red-teaming processes to guard against abuse

Technology platform developers and AI researchers should assume that disinformation actors will misuse the platform features and capabilities of their research. Threat modeling is a cybersecurity practice that helps proactively identify areas ripe for adversary exploitation. Through this process, developers map risks and vulnerabilities of new features and capabilities, which helps to anticipate new threat tactics and identify potential mitigations. Red teaming, a practice of emulating a threat actor and attacking systems from an adversary's point of view, helps determine if avenues for misuse are mitigated.

Technology platforms and AI researchers should adapt these cybersecurity practices to identify how disinformation actors may misuse their platform features and AI research. They should staff threat modeling teams and integrate them into the product development process. AI developers that are commercializing content generation capabilities or offering "AI-as-a-Service" should use threat modeling to establish vetting procedures, access controls, and threat hunting processes to monitor for misuse. In addition, they should incorporate red-teaming techniques to emulate disinformation operators. Fostering partnerships with researchers akin to the ethical hacking community can help identify cross-platform threats. Developing a process, a safe environment, and a pipeline of external red teams can help test the platforms and AI-enabled systems for misuse and proactively identify opportunities for mitigations.

Build and apply ethical principles for the publication of AI research that can fuel disinformation campaigns

The open nature of AI research is critical to innovation and advancement of U.S. leadership in this field, and overregulating this ecosystem could have an undesirable effect of slowing down innovation. The ML research field is also in a "pre-Hippocratic oath era" as it continues to debate the value of releasing new capabilities openly weighed against potential harm.[209] While open publication of models, code, and tutorials advances research, it also helps threat actors. Fine-tuning ML models for tailored tasks can shorten the development time and reduce computing costs making it more likely that threat actors will exploit these freely accessible capabilities.

Prominent voices in the AI research community have called for greater responsibility over publishing models and research findings, but these cases are an exception, not the rule. Some researchers have limited the scope of their releases due to malicious use concerns.[210] Setting an example for the broader field, organizers of a prominent AI conference now require submitted papers to consider how their research may be misused.[211] Other researchers openly released AI-enabled systems with implications for disinformation operations, but documented the likely abuse scenarios.[212] Still others advocate for open release of research in order to crowdsource mitigations and plan to make powerful large language models publicly available.[213] They argue that as the barriers to develop generative capabilities around the world lower, fighting this diffusion by limiting access is futile.

The AI research community should assume that disinformation operators will misuse their openly released research and develop a publication risk framework to guard against it. Researchers should weigh national security and societal harm in decisions about how widely to publicize findings, and they should build mitigations prior to release. [214] The "do no harm" principle should weigh heavily in the decision of how much of the model, code, and tutorial to release publicly.[215]

Establish a process for the media to report on disinformation without amplifying it

Print and broadcast media organizations may be both the targets and unwitting spreaders of disinformation. Some media organizations have established processes to guard against inadvertently amplifying disinformation campaigns, but the practice is not universal. National media organizations should use a threat modeling approach to examine their processes and understand how the flow of information to them and through them can be exploited by disinformation operators for amplification, particularly around national political events such as elections.[216] Threat modeling and table-top exercises can help develop procedures for how a newsroom might report on a potential hack-forge-leak operation, particularly as ML-enhanced forgeries may become more common. It can identify gaps in capacity and forensic capabilities to determine the provenance of content that is generated or modified by ML and help direct limited resources.[217] Researchers and information security professionals can combine forces with journalists to develop a threat modeling tool kit for small and medium-sized media organizations that do not have the capacity to hire threat modeling teams, yet play a critical role in local communities. This practice can help build resilience against disinformation campaigns, whether AI-powered or not.

Reform recommendation algorithms

Social media recommendation algorithms can contribute to information bubbles, polarization, and radicalization through algorithmic deterioration that reinforces users' choices. Malicious actors can also exploit the algorithms to increase exposure to disinformation. The recently reinvigorated debate about how algorithms should prioritize information on users' social media feeds has so far yielded few actionable or satisfying solutions.

While consensus on the need to reform recommendation algorithms is emerging, there is no consensus on how to do so. Some propose to minimize personalization of newsfeeds and present information to users chronologically. Other approaches may involve giving users a way to reset the assumptions the

algorithm has made about them based on their past actions and choices of similar users, or their look-alikes. Platforms and AI researchers should invest in measures to increase transparency such as increasing the ability of algorithms to explain their choices. Other research should explore how these systems can be rebalanced to counter the filter-bubble effect. Solutions should empower users to make informed decisions by exposing origins of content. Platforms should enable independent audits by vetted researchers to help illuminate how recommendation systems should change in order to fulfill their original purpose—to connect and inform human societies, not divide and disinform them.

Raise awareness and build public resilience against ML-enabled disinformation

Humans are the ultimate line of defense against disinformation campaigns. While most countermeasures focus on the supply of disinformation, the demand can fuel its spread. The proliferation of health-related disinformation in 2020—an "infodemic"—spurred the discussion about how a public health approach of inoculating the population can apply to preventing the spread of disinformation about COVID-19 and beyond.[218] Nations on the frontlines of disinformation operations have successfully implemented programs to educate their populations to discern online manipulation.[219] A variety of tools from the civil society have emerged to help raise digital literacy, yet few account for the potential of ML-enabled disinformation.[220] The U.S. government, the private sector, and state and local governments should take measures to build societal resilience against disinformation campaigns through school and adult education programs. Social media platforms working with civil society can identify the most targeted communities and arm them with tools to help discern ML-enabled disinformation techniques. The United States should incorporate lessons learned from successful digital media literacy programs around the world, many of which it supported and funded, to build resilience to disinformation campaigns at home.

Conclusion

Disinformation operations have become a global phenomenon. A race to AI dominance is underway, motivated by national prestige and strategic advantage. State and non-state actors can now augment their disinformation operations with ML-enabled capabilities more readily as the remaining barriers fall. Open-source research is the norm, and computing costs continue to decrease.

Disinformation campaigns have already succeeded beyond the expectations of their creators. A natural next evolution is the application of AI capabilities to these campaigns when such techniques increase effectiveness and scale. Natural language processing, sentiment analysis and social listening applications are already likely harnessing voluminous digital footprints and giving threat actors insight into fissures to widen, informational ambiguities to exploit, and individuals to mobilize. On the horizon, generative models are likely to significantly scale up the creation of engaging content and challenge efforts to distinguish the authentic from the synthetic. Natural language generation is steadily improving and becoming available in more languages. Just beyond the horizon, conversational AI systems may become a new vector of tailored and personalized disinformation to radicalize and mobilize humans into action or inaction.

Beyond tactical applications, our findings hint at the underlying challenges in the current information ecosystem that AI/ML technologies are likely to exacerbate: the cross-platform nature of modern disinformation campaigns, the targeting of traditional broadcast media to super-spread their messages, and the role of transnational actors, such as "influence as a service" companies, in polluting public discourse. Chief among the challenges is the lack of public consensus on several unresolved questions primed for public debate. Does First Amendment protection extend to artificially generated or amplified speech? Where should societies draw the line at which AI-generated falsehoods cause harm and lose their free speech protection? [221] What are the rights and principles to guide how citizens engage in the AI/ML-enabled world?[222] The rapid advancement of AI/ML capabilities make this debate more urgent than ever.

We offer select recommendations that can help build defenses and better protect users from disinformation efforts. Our work illuminates the plausible direction of what some researchers call "information disorder," [223] but we recognize that far more work remains to be done. Our report focuses on social media and the online information environment because they will be primarily impacted by AI-enabled capabilities. Yet they are part of a larger challenge to the epistemic security of democratic societies. As a dual-use technology, AI/ML may also powerfully enhance defenses against disinformation operations, although we do not assess their effectiveness in this report. Further studies should illuminate how AI/ML can help disrupt operations at each stage of a disinformation campaign. Finally, as "AI-as-a-Service" firms commercialize advancing capabilities, such as generative models and conversational AI, there is a need for greater examination on how best to safeguard these systems and build joint defense-in-depth of the broader information environment.

We hesitate to forecast where this evolution leads. It is possible that more people will turn away from the digital environment, particularly if accompanied by growing resilience, education, and degree of skepticism to discern fact from fiction. Yet it is also plausible that the normalization of AI-generated disinformation may create a downward spiral in which a more cynical audience provides fertile ground for even more false information. As democratic societies build resilience and raise awareness about artificially amplified speech, they may struggle to maintain healthy skepticism without succumbing to cynicism and distrust.

It is also likely that, if left unchecked, AI/ML-enabled disinformation operations will deepen existing societal fissures and complicate the task of sustaining a shared public square that is critical for democracies to function.

Artificial intelligence offers enormous promise to advance human progress as well as powerful capabilities to disrupt it. A critical question for the world's democracies is how to harness the promise of AI/ML and mitigate its potential harms without betraying their foundational principles.

Authors

Katerina Sedova is a research fellow with the CyberAI team at the Center for Security and Emerging Technology. Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan are former student research analysts.

Acknowledgments

For their feedback and insights, we thank our CSET colleagues: John Bansemer, Ben Buchanan, Margarita Konaev, Andrew Imbrie, Micah Musser, Dakota Cary, and Andrew Lohn. We are grateful for the thoughtful reviews provided by Miriam Matthews of the RAND Corporation and Alicia Wanless of the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace. Finally, we thank Hannah Stone, Melissa Deng, Shelton Fitch, and Lynne Weil for editorial support. Any errors that remain are the fault of the authors alone.

Document Identifier: doi: 10.51593/2021CA011

Endnotes

[1] Select Committee on Intelligence, *Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, *United States Senate*, Volume 2, November 10, 2020, https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf.

[2] Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework" (Center for Security and Emerging Technology, forthcoming/December 2021).

[3] Author's interview (conducted on Dec 12, 2016) with Dr. Andrew Kuchins, then senior fellow and research professor at the Center for Eurasian, Russian, and East European Studies (CERES) at Georgetown University's Walsh School of Foreign Service. He is also a senior associate (non-resident) of the Center for Strategic and International Studies' (CSIS) Russia and Eurasia Program, which he led from 2007 to 2015. From 2000 to 2006, Kuchins was a senior associate at the Carnegie Endowment for International Peace, where he previously served as director of its Russian and Eurasian Program in Washington, DC, from 2000 to 2003 and again in 2006.

[4] Daniel Treisman, "Why Putin Took Crimea: The Gambler in the Kremlin," *Foreign Affairs* 95, no. 3 (May/June 2016): 50.

[5] Anton Bebler, "The Russian-Ukrainian Conflict Over Crimea," *TEORIJA IN PRAKSA* let. 52, 1–2 (2015): 203.

[6] Bebler, "The Russian-Ukrainian Conflict Over Crimea," 203.

[7] Elizabeth A. Wood, "Chronology: The War for Crimea and Ukraine," in *Roots of Russia's War in Ukraine* (Woodrow Wilson Center Press/Columbia University Press, 2016).

[8] Fiona Hill and Clifford G. Gaddy, *Mr. Putin: Operative in the Kremlin (Geopolitics in the 21st Century)* (Washington, DC: Brookings Institution Press, 2013), ch. 1.

[9] "Dezinformatsiya" is a Russian word, defined in the Great Soviet Encyclopedia as the "dissemination of misleading or false information, used as a method of political propaganda aimed to mislead public opinion," *Great Soviet Encyclopedia Online*, https://bse.slovaronline.com/10240-DEZINFORMATSIYA. See also Select Committee on Intelligence, *Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, Volume 2.

[10] Select Committee on Intelligence, *Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, Volume 2, 11; Alicia Wanless and James Pamment, "How Do You Define a Problem Like Influence?," *Journal of*

Information Warfare 18, no. 3 (Winter 2019), December 30, 2019, https://carnegieendowment.org/2019/12/30/how-do-you-define-problem-like-influence-pub-80716.

[11] Jon Bateman et al., "Measuring the Effects of Influence Operations: Key Findings and Gaps From Empirical Research" (Carnegie Endowment for International Peace, June 28, 2021), https://carnegieendowment.org/2021/06/28/measuring-effects-of-influence-operations-key-findings-and-gaps-from-empirical-research-pub-84824.

[12] Nathaniel Gleicher et al., "The State of Influence Operations 2017-2020" (Facebook, May 2021), https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf.

[13] For China's foray into international influence campaigns, see: Twitter Safety, "Disclosing Networks of State-linked Information Operations We've Removed," *Twitter*, June 12, 2020, https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html; Jacob Wallis et al., "Retweeting Through the Great Firewall" (Australian Strategic Policy Institute, June 12, 2020), https://www.aspi.org.au/report/retweeting-through-great-firewall; Carly Miller, Vanessa Molter, Isabella Garcia-Camargo, and Renée DiResta, "Sockpuppets Spin COVID Yarns: An Analysis of PRC-Attributed June 2020 Twitter Takedown" (Stanford Internet Observatory, June 17, 2020), https://stanford.app.box.com/v/sio-twitter-prc-june-2020; Ryan Serabian and Lee Foster, "Pro-PRC Influence Campaign Expands to Dozens of Social Media Platforms, Websites, and Forums in at Least Seven Languages, Attempted to Physically Mobilize Protesters in the U.S," *Mandiant Threat Research*, September 8, 2021, https://www.mandiant.com/resources/pro-prc-influence-campaign-expands-dozens-social-media-platforms-websites-and-forums. For Iran's evolution: Ben Nimmo et al., "Iran's Broadcaster: Inauthentic Behavior" (Graphika, May 2020), https://public-assets.graphika.com/reports/graphika_report_irib_takedown.pdf; See also: Alice Revelli and Lee Foster, "'Distinguished Impersonator' Information Operation That Previously Impersonated U.S. Politicians and Journalists on Social Media Leverages Fabricated U.S. Liberal Personas to Promote Iranian Interests," *Mandiant*, February 12, 2020, https://www.fireeye.com/blog/threat-research/2020/02/information-operations-fabricated-personas-to-promote-iranian-interests.html. See also: Mona Elswah, Phillip N. Howard, and Vidya Narayanan, "Iranian Digital Interference in the Arab World" (Oxford Internet Institute, April 3, 2019), https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/04/Iran-Memo.pdf. See also: FireEye Intelligence, "Suspected Iranian Influence Operation Leverages Network of Inauthentic News Sites & Social Media Targeting Audience in the U.S., UK, Latin America, Middle East," August 21, 2018, Mandiant, https://www.fireeye.com/blog/threat-research/2018/08/suspected-iranian-influence-operation.html; William Evanina, "100 Days Until Election 2020," Office of the Director of National Intelligence,

July 24, 2020, https://www.dni.gov/index.php/newsroom/press-releases/item/2135-statement-by-ncsc-director-william-evanina-100-days-until-election-2020.

[14] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation" (Oxford Internet Institute, 2020), https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report-2020-v.2.pdf; Diego A. Martin, Jacob N. Shapiro, and Julia Ilhardt, "Trends in Online Influence Efforts" (Empirical Studies of Conflict Project, Princeton University, 2020), https://esoc.princeton.edu/publications/trends-online-influence-efforts.

[15] Daniel Kahneman, *Thinking, Fast and Slow* (New York, NY: Farrar, Straus and Giroux, 2011).

[16] Margarita Konaev et al., "Headline or Trend Line? Evaluating Chinese-Russian Collaboration in AI" (Center for Security and Emerging Technology, August 2021), https://cset.georgetown.edu/publication/headline-or-trend-line/.

[17] National Security Commission on Artificial Intelligence, *Final Report* (Washington, DC: NSCAI, March 1, 2021), https://reports.nscai.gov/final-report/chapter-1/.

[18] Minority Media, Homeland Security & Governmental Affairs Committee, "Tech Leaders Support Portman's Bipartisan Deepfake Task Force Act to Create Task Force at DHS to Combat Deepfakes," United States Senate, July 30, 2021, https://www.hsgac.senate.gov/media/minority-media/tech-leaders-support-portmans-bipartisan-deepfake-task-force-act-to-create-task-force-at-dhs-to-combat-deepfakes; Cristiano Lima, "The Technology 202: As senators zero in on deepfakes, some experts fear their focus is misplaced," *The Washington Post*, August 6, 2021, https://www.washingtonpost.com/politics/2021/08/06/technology-202-senators-zero-deepfakes-some-experts-fear-their-focus-is-misplaced/.

[19] Elizabeth Seger et al., "Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world" (Center for the Study of Existential Risk, University of Cambridge, October 14, 2020), https://www.cser.ac.uk/resources/epistemic-security/.

[20] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[21] "Glossary," Center for Security and Emerging Technology.

[22] Greg Allen, *Understanding AI Technology* (Washington, DC: Joint Artificial Intelligence Center, United States Department of Defense, April 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf; IBM Cloud Education, "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks:

What's the Difference?," *IBM*, May 27, 2020, https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

[23] Twint Project, https://github.com/twintproject/twint.

[24] CloudFlare Learning Center, "What is Data Scraping?," CloudFlare, https://www.cloudflare.com/learning/bots/what-is-data-scraping/.

[25] Steve Lohr, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," *The New York Times*, August 17, 2014, https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

[26] Tim Stobierski, "Data Wrangling: What It Is & Why It's Important," Harvard Business School, January 19, 2021, https://online.hbs.edu/blog/post/data-wrangling.

[27] Jonathan Vanian, "The A.I. Boom Helped This Data Cleaning Startup Collect $100 Million From Investors," *Fortune*, September 12, 2019, https://fortune.com/2019/09/12/data-cleaning-startup-investors/.

[28] IBM Cloud Education, "What is Natural Language Processing?," IBM, July 2, 2020, https://www.ibm.com/cloud/learn/natural-language-processing.

[29] IBM Cloud Education, "What is Natural Language Processing?"; Allen, *Understanding AI Technology*.

[30] Tom B. Brown et al., "Language Models are Few-Shot Learners," *arXiv* preprint arXiv:2005.14165 (2020), https://arxiv.org/abs/2005.14165; Kyle Wiggers, "Huawei trained the Chinese-language equivalent of GPT-3," *VentureBeat*, April 29, 2021, https://venturebeat.com/2021/04/29/huawei-trained-the-chinese-language-equivalent-of-gpt-3/; Wei Zeng et al., "PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation," arXiv preprint arXiv:2104.12369 (2021), https://arxiv.org/pdf/2104.12369.pdf.

[31] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova, "Truth, Lies, and Automation: How Language Models Could Change Disinformation" (Center for Security and Emerging Technology, May 2021), https://cset.georgetown.edu/publication/truth-lies-and-automation/; "GPT-3 and PanGu Alpha by Huawei," *The Silicon Trend*, May 7, 2021, https://thesilicontrend.com/pangu-alpha-a-chinese-equivalent-of-gpt-3.

[32] John Seabrook, "The Next Word: Where Will Predictive Text Take Us?," *The New Yorker*, October 14, 2019, https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-

write-for-the-new-yorker; Buchanan, Lohn, Musser, and Sedova, ""Truth, Lies, and Automation."

[33] Yaser Martinez Palenzuela, "Awesome GPT-3," *GitHub*, https://github.com/elyase/awesome-gpt3/; GPT-3, "A robot wrote this entire article. Are you scared yet, human?," *The Guardian*, September 8, 2020, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3; Brown et al., "Language models are few-shot learners'; Mrinal Mohit (@wowitsmrinal), "Tired: Making your own memes, Wired: Asking @OpenAI's #gpt3 to make memes. Amazed to see how much of cultural subtext and nuance language models can pick up on," Twitter, July 25, 2020, https://twitter.com/wowitsmrinal/status/1287175391040290816?s=20.

[34] OpenAI Beta, "OpenAI Technology, Just an HTTPS Call Away," OpenAI, https://beta.openai.com/; Sid Bharath (@Siddharth87), "I'm now playing around with writing ads on Google. I've fed the AI the top ad copy for 'sales engagement software'" and it generated two really useful outputs below the dotted line that I could run without any edits," Twitter, July 13, 2020, https://twitter.com/Siddharth87/status/1282823360825581568.

[35] Public Press Release, "NAVER unveils HyperCLOVA, Korea's first super-sized AI . . . 'Leading the age of AI for all,'" Naver Corp, May 25, 2021, https://www.navercorp.com/promotion/pressReleasesView/30546; SberAI, "Russian GPT3 Models," *Sberbank AI Github Repository*, https://github.com/sberbank-ai/ru-gpts#readme; "PAGnol," LightOnAI, https://lair.lighton.ai/pagnol/; "Announcing AI21 Studio and Jurassic-1 Language Models," AI21Labs, August 11, 2021, https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1; Ali Alvi and Paresh Kharya, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model - Microsoft Research," *Microsoft Research Blog*, October 11, 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge, MA: *MIT Press*, 2016), 534, https://www.deeplearningbook.org/contents/representation.html.

[37] Goodfellow, Bengio, and Courville, *Deep Learning*.

[38] Gleicher et al., "The State of Influence Operations 2017-2020."

[39] Nicholas Confessore, Gabriel J.X Dance, Richard Harris, and Mark Hansen, "The Follower Factory," *The New York Times*, January 27, 2018, https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

[40] Steven L. Scott and Hal R. Varian, "Predicting the Present with Bayesian Structural Time Series," June 28, 2013, SSRN, http://dx.doi.org/10.2139/ssrn.2304426.

[41] Madhusree Mukerjee, "How Fake News Goes Viral – Here's the Math," *Scientific American*, July 14, 2017, https://www.scientificamerican.com/article/how-fake-news-goes-viral-mdash-heres-the-math/.

[42]Richard Colbaugh and Kristin Glass, "Leveraging sociological models for prediction I: Inferring adversarial relationships," *2012 IEEE International Conference on Intelligence and Security Informatics*, 2012, 66-71, https://doi.org/10.1109/ISI.2012.6284093; Richard Colbaugh and Kristin Glass, "Leveraging sociological models for prediction II: Early warning for complex contagions," *2012 IEEE International Conference on Intelligence and Security Informatics*, 2012, 72-77, https://doi.org/10.1109/ISI.2012.6284094.

[43] Google Search, "How Search Organizes Information," *Google*, https://www.google.com/search/howsearchworks/crawling-indexing/.

[44] "Google Alerts," *Google*, https://www.google.com/alerts.

[45] Tony Tran, "What is Social Listening, Why it Matters, and 10 Tools to Make it Easier," *Hootsuite Blog*, March 3, 2020, https://blog.hootsuite.com/social-listening-business/; Werner Geyser, "Top 21 Social Media Listening Tools for 2021," *Influencer Marketing Hub*, September 9, 2021, https://influencermarketinghub.com/social-media-listening-tools; Mention, https://mention.com/en/listen/; "YouScan Review," Influencer Marketing Hub, https://influencermarketinghub.com/youscan/.

[46] Jason Wilcox, "The Multi-Dimensional Value of Public Twitter Data for Real-Time Event Detection," *Dataminr*, December 6, 2019, https://www.dataminr.com/blog/the-multi-dimensional-value-of-public-twitter-data-for-real-time-event-detection; "Platform Overview," *IDenTV*, http://www.identv.com/platform.html.

[47] Sebastian Ruder, "NLP-Progress: Sentiment Analysis," *GitHub*, https://github.com/sebastianruder/NLP-progress/blob/master/english/sentiment_analysis.md; Zhilin Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding," arXiv preprint arXiv:1906.08237 (2020), https://arxiv.org/pdf/1906.08237.pdf.

[48] Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in *IEEE Access* 7 (2019): 44883-44893, https://doi.org/10.1109/ACCESS.2019.2909180.

[49] For one example of such tools, see "Watson Natural Language Understanding," IBM, https://www.ibm.com/cloud/watson-natural-language-understanding.

[50] Vincenzo Ciancaglini et al., "Malicious Uses and Abuses of Artificial Intelligence" (Trend Micro Research, 2020), https://documents.trendmicro.com/assets/white_papers/wp-malicious-uses-and-abuses-of-artificial-intelligence.pdf; Singularex, "The Black Market for Social Media Manipulation" (NATO Strategic Communications Centre of Excellence, January 19, 2019), https://stratcomcoe.org/publications/the-black-market-for-social-media-manipulation/103; "Natural Language Toolkit: 3.6.3 Documentation," NLTK Project, https://www.nltk.org/; "Natural Language Processing in Bot Automation," *Blackhatworld Forum*, August 29, 2018, http://webcache.googleusercontent.com/search?q=cache:cuh7cESvDlAJ:https://www.blackhatworld[...]n-bot-automation.1052772/&hl=en&gl=us&strip=1&vwsrc=0.

[51] Abeer ALDayel and Walid Magdy, "Stance Detection on Social Media: State of the Art and Trends," *Information Processing & Management* 58, no. 4 (2021), https://doi.org/10.1016/j.ipm.2021.102597.

[52] Dilek Küçük and Fazli Can, "Stance Detection: A Survey," *ACM Computing Surveys* 53, no. 1 (May 2020): Article 12, https://doi.org/10.1145/3369026.

[53] Mirko Lai et al., "Multilingual Stance Detection in Social Media Political Debates," *Computer Speech & Language* 63 (2020), https://doi.org/10.1016/j.csl.2020.101075.

[54] "*ORA-LITE Project," Center for Computational Analysis of Social and Organizational Systems, Carnegie Mellon University, http://www.casos.cs.cmu.edu/projects/ora/; Kathleen M. Carley, "NetMapper for Extracting Networks from Texts Has Been Released," *Netanomics*, January 4, 2017, https://netanomics.com/netmapper-for-extracting-networks-from-texts-has-been-released/.

[55] "VoyagerCheck Platform," Voyager Labs, https://voyagerlabs.co/platforms/voyagercheck.

[56] Yuxin Ding et al., "Predicting the Attributes of Social Network Users Using a Graph-Based Machine Learning Method," *Computer Communications* 73 (2016): Part A, https://doi.org/10.1016/j.comcom.2015.07.007.

[57] "Trendscope," *Black Swan Data*, https://www.blackswan.com/trendscope/; Andrea Vattani, "Understanding at Scale," Spiketrap, https://www.spiketrap.io/science/

[58] "Discover Your Marketing Personas with Audience Segmentation," Socialbakers, https://www.socialbakers.com/feature/audience-segmentation

[59] Facebook for Developers, "Lookalike Audiences," *Facebook*, https://developers.facebook.com/docs/marketing-api/audiences/guides/lookalike-audiences.

[60] "About Facebook Ads," *Facebook*, https://www.facebook.com/ads/about/?entry_product=ad_preferences; Caitlin Dewey, "98 Personal Data Points that Facebook Uses to Target Ads to You," *The Washington Post*, August 19, 2016, https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/; Geoffrey A. Fowler, "Facebook Will Now Show You Exactly Now It Stalks You — Even When You're Not Using Facebook," *The Washington Post*, January 28, 2020, https://www.washingtonpost.com/technology/2020/01/28/off-facebook-activity-page/; Facebook Policy, "Data Policy," *Facebook*, https://www.facebook.com/policy.php.

[61] "Psychographics," Dictionary of Psychology, American Psychological Association, https://dictionary.apa.org/psychographics; "What is Psychographics? Understanding the Tech That Threatens Elections" (CBInsights, 2020), https://www.cbinsights.com/reports/CB-Insights-What-is-Psychographics.pdf.

[62] "Big 5 Personality Traits," *Psychology Today*, https://www.psychologytoday.com/us/basics/big-5-personality-traits; "AI-Powered Consumer Research & Intelligence," Resonate, https://www.resonate.com/resonate-ignite-platform/.

[63] "Personality Insights: The Science Behind the Service," *IBM*, August 18, 2021, https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-science; H. Andrew Schwartz et al., "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS ONE*, September 25, 2013, https://doi.org/10.1371/journal.pone.0073791.

[64] Geoffrey A. Fowler, "How Political Campaigns Get Your Phone Number and Other Data," *The Washington Post*, October 27, 2020, https://www.washingtonpost.com/technology/2020/10/27/political-campaign-data-targeting; Douglas MacMillan, "Vermont's New Data Broker Law Shows Challenge of Regulating Companies Compiling and Selling Your Personal Information," *The Washington Post*, June 24, 2019, https://www.washingtonpost.com/business/2019/06/24/data-brokers-are-getting-rich-by-selling-your-secrets-how-states-are-trying-stop-them/.

[65] United States Federal Trade Commission, "Equifax Data Breach Settlement," January 2020, https://www.ftc.gov/enforcement/cases-proceedings/refunds/equifax-data-breach-settlement; Michael Hill and Dan

Swinhoe, "The 15 Biggest Data Breaches of the 21st Century," *CSO Online*, July 16, 2021, https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html; Zak Doffman, "New Data Breach Has Exposed Millions of Fingerprint and Facial Recognition Records: Report," *Forbes*, August 14, 2019, https://www.forbes.com/sites/zakdoffman/2019/08/14/new-data-breach-has-exposed-millions-of-fingerprint-and-facial-recognition-records-report/; Brendan Koerner, "Inside the OPM Hack, The Cyberattack that Shocked the US Government," *WIRED*, October 23, 2016, https://www.wired.com/2016/10/inside-cyberattack-shocked-us-government/.

[66] "This X Does Not Exist," https://thisxdoesnotexist.com/.

[67] Davey Alba, "Facebook Discovers Fakes That Show Evolution of Disinformation," *The New York Times*, December 20, 2019, https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html.

[68] Raphael Satter, "Experts: Spy Used AI-generated Face to Connect with Targets," AP News, June 13, 2019, https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

[69] Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior From Georgia, Vietnam and the US," *Facebook Newsroom*, December 20, 2019, https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/; Camille François and Iain Robertson, editors, "#Operation FFS: Fake Face Swarm" (Graphika and DFRLab, December 2019), https://public-assets.graphika.com/reports/graphika_report_operation_ffs_fake_face_storm.pdf.

[70] Ben Nimmo, Camille Francois, C. Shawn Eib, and Lea Ronzaud, "Spamouflage Goes to America" (Graphika, August 2020), https://graphika.com/reports/spamouflage-dragon-goes-to-america/; William Evanina, "Election Threat Update for the American Public," Office of the Director of National Intelligence, August 7, 2020, https://www.dni.gov/index.php/newsroom/press-releases/item/2139-statement-by-ncsc-director-william-evanina-election-threat-update-for-the-american-public.

[71] Ben Nimmo, Camille Francois, C. Shawn Eib, and Lea Ronzaud, "IRA Again: Unlucky Thirteen" (Graphika, September 2020), https://public-assets.graphika.com/reports/graphika_report_ira_again_unlucky_thirteen.pdf; "August 2020 Coordinated Inauthentic Behavior Report," *Facebook*, August 2020, https://about.fb.com/news/2020/09/august-2020-cib-report/; Twitter Safety (@TwitterSafety), "We suspended five Twitter accounts for platform manipulation that we can reliably attribute to Russian state actors. As standard, they will be included in updates to our database of information operations in the

coming weeks to empower academic research," Twitter, September 1, 2020, https://twitter.com/TwitterSafety/status/1300848632120242181?s=20.

[72] Graphika Team, "Step into My Parler: Suspected Russian Operation Targeted Far-Right American Users on Platforms Including Gab and Parler, Resembled Recent IRA-Linked Operation that Targeted Progressives" (Graphika, October 2020), https://graphika.com/reports/step-into-my-parler/.

[73] Graphika Team and Stanford Internet Observatory, "More Troll Kombat: French and Russian Influence Operations Go Head to Head Targeting Audiences in Africa" (Graphika and Stanford Internet Observatory, December 15, 2020), https://publicassets.graphika.com/reports/graphica_stanford_report_more_troll_kombat.pdf; Kimberly Marten, "The GRU, Yevgeny Prigozhin, and Russia's Wagner Group: Malign Russian Actors and Possible U.S. Responses," Testimony before Committee on Foreign Affairs, Subcommittee on Europe, Eurasia, Energy, and the Environment, United States House of Representatives, July 7, 2020, https://www.congress.gov/116/meeting/house//110854/witnesses/HHRG-116-FA14-Wstate-MartenK-20200707.pdfFA14-Wstate-MartenK-20200707.pdf; Graphika Team, "Fake Cluster Boosts Huawei" (Graphika, January 2021), https://publicassets.graphika.com/reports/graphika_report_fake_cluster_boosts_huawei.pdf

[74] Philip Tully and Lee Foster, "Repurposing Neural Networks to Generate Synthetic Media for Information Operations," *Mandiant*, August 5, 2020, https://www.fireeye.com/blog/threat-research/2020/08/repurposing-neural-networks-to-generate-synthetic-media-for-information-operations.html.

[75] Tao Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," https://openaccess.thecvf.com/content_cvpr_2018/papers/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.pdf.

[76] Shiv (@Shivkantb), "Introducing GPT3 x GAN AI generated faces using natural language powered by GPT-3. Eg – 'Generate a female face with blonde hair and green eyes,'" Twitter, September 4, 2020, https://twitter.com/shivkanthb/status/1302039696692789249?s=20.

[77] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[78] Vlad Alex (Merzmensch), "This [item] Does Not Exist," *Towards Data Science*, December 25, 2020, https://towardsdatascience.com/this-item-does-not-exist-2defbac76b39; Aggregate Intellect, "Awesome-Does-Not-Exist," *Github*, March 27, 2019, https://github.com/Aggregate-Intellect/awesome-does-not-exist.

[79] "Analysis of an October 2020 Facebook Takedown Linked to U.S. Political Consultancy Rally Forge," *Stanford Internet Observatory Blog*, October 8, 2020, https://cyber.fsi.stanford.edu/news/oct-2020-fb-rally-forge.

[80] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova, "Truth, Lies, and Automation: How Language Models Could Change Disinformation" (Center for Security and Emerging Technology, May 2021), https://cset.georgetown.edu/publication/truth-lies-and-automation/.

[81] Cici Zhang, "Baidu's AI Produces Short Videos in One Click," *IEEE Spectrum*, May 20, 2020, https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/baidus-ai-produces-short-videos-in-one-click.

[82] Paul Lambert, "Subject: Write Emails Faster with Smart Compose In Gmail," *Google*, May 8, 2018, https://www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail/; John Seabrook, "The Next Word: Where Will Predictive Text Take Us?," *The New Yorker*, October 14, 2019, https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker.

[83] Nimmo et al., "IRA Again: Unlucky Thirteen"; Ben Nimmo (@benimmo), "One more thought on PeaceData: the thing with co-opting real people is that it creates a lot of witnesses," *Twitter*, September 4, 2020, https://twitter.com/benimmo/status/1301950111916883968?s=20.

[84] Andrej Karpathy et al., "Generative Models," *OpenAI Blog*, June 16, 2016, https://openai.com/blog/generative-models/; Brad D. Williams, "Researchers Warn Of 'Dangerous' Artificial Intelligence-Generated Disinformation At Scale," *Breaking Defense*, September 30, 2021, https://breakingdefense.com/2021/09/researchers-warn-of-dangerous-artificial-intelligence-generated-disinformation-at-scale/amp/.

[85] "News and Announcements: GPT-J-6B," *EleutherAI Blog*, June 8, 2021, https://www.eleuther.ai/.

[86] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[87] Anton Troianovski, "A former Russian troll speaks: 'It was like being in Orwell's world,'" *The Washington Post*, February 17, 2018, https://www.washingtonpost.com/news/worldviews/wp/2018/02/17/a-former-russian-troll-speaks-it-was-like-being-in-orwells-world/; Aric Toler, "Inside the Kremlin Troll Army Machine: Templates, Guidelines, and Paid Posts," *Global Voices*, March 14, 2015, https://globalvoices.org/2015/03/14/russia-kremlin-troll-army-examples/.

[88] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York, NY: Farrar, Straus, and Giroux: New York, 2020);

Polina Rusyaeva and Andrei Zaharov, "How 'Troll Factory' worked the U.S. Elections," *RBK Magazine*, October 17, 2017, https://web.archive.org/web/20210303095306/https://www.rbc.ru/magazine/2017/11/59e0c17d9a79470e05a9e6c1.

[89] Kris McGuffie and Alex Newhouse, "The Radicalization Risks Posed by GPT-3 and Other Advanced Neural Language Models," Middlebury Institute of International Studies at Monterey, September 14, 2020, https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language.

[90] Buchanan et al., "Truth, Lies, and Automation."

[91] Sarah Kreps, Miles McCain, and Miles Brundage, "All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *SSRN*, September 24, 2020, http://dx.doi.org/10.2139/ssrn.3525002.

[92] For a conceptualization of Hack-Forge-Leak campaigns, see Rid, *Active Measures*.

[93] Ben Nimmo, "UK Trade Leaks and Secondary Infektion: New Findings and Insights from a Known Russian Operation" (Graphika, December 2019), https://public-assets.graphika.com/reports/graphika_report_uk_trade_leaks_&_secondary_infektion.pdf; Lee Foster et al., "Ghostwriter Update: Cyber Espionage Group UNC1151 Likely Conducts Ghostwriter Influence Activity," *Mandiant*, April 28, 2021, https://www.fireeye.com/blog/threat-research/2021/04/espionage-group-unc1151-likely-conducts-ghostwriter-influence-activity.html.

[94] Tom Simonite, "Give These Apps Some Notes and They'll Write Emails for You," *WIRED*, October 18, 2020, https://www.wired.com/story/give-apps-notes-they-write-emails/.

[95] Sberbank AI, "Ru-GPTs: Russian GPT-3 Models," *GitHub*, August 17, 2020, https://github.com/sberbank-ai/ru-gpts#readme; "NAVER unveils HyperCLOVA, Korea's first super-sized AI . . . ''Leading the age of AI for all.'"

[96] Connor Leahy, "Why Release a Large Language Model?" *EleutherAI Blog*, June 2, 2021, https://blog.eleuther.ai/why-release-a-large-language-model.

[97] Abhishek Iyer, "Stuck in GPT-3's waitlist? Try out the AI21 Jurassic-1," *VentureBeat*, September 11, 2021, https://venturebeat.com/2021/09/11/stuck-in-gpt-3s-waitlist-try-out-the-ai21-jurassic-1/.

[98] Zhang, "Baidu's AI Produces Short Videos in One Click."

[99] Tao Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks."

[100] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Scott Gray, "DALL·E: Creating Images from Text," *OpenAI Blog*, January 5, 2021, https://openai.com/blog/dall-e/.

[101] Alex Zhavoronkov, "Wu Dao 2.0 - Bigger, Stronger, Faster AI From China," *Forbes*, July 19, 2021, https://www.forbes.com/sites/alexzhavoronkov/2021/07/19/wu-dao-20bigger-stronger-faster-ai-from-china/.

[102] Abel L. Peirson and E. Meltem Tolunay, "Dank learning: Generating memes using deep neural networks," arXiv preprint arXiv:1806.04510 (2018), https://arxiv.org/abs/1806.04510.

[103] Martins Frolovs, "Teaching GPT-2 Transoformer a Sense of Humor," *Towards Data Science*, December 17, 2019, https://towardsdatascience.com/teaching-gpt-2-a-sense-of-humor-fine-tuning-large-transformer-models-on-a-single-gpu-in-pytorch-59e8cec40912.

[104] Karen Hao, "Deepfake Porn is Ruining Women's Lives. Now the Law May Finally Ban It," *MIT Technology Review*, February 12, 2021, https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/; Sensity Team, "Automating Image Abuse: Deepfake Bots on Telegram," *SensityAI Blog*, October 20, 2020, https://sensity.ai/blog/deepfake-detection/automating-image-abuse-deepfake-bots-on-telegram/; Ciancaglini et al., "Malicious Uses and Abuses of Artificial Intelligence"; Ingrid Lunden, "Snapchat Quietly Acquired AI Factory, the Company Behind Its New Cameos Feature, for $166M," *TechCrunch*, January 3, 2020, https://techcrunch.com/2020/01/03/snapchat-quietly-acquired-ai-factory-the-company-behind-its-new-cameos-feature-for-166m/; Jeremy Kahn, "These Deepfake Videos of Putin and Kim Have Gone Viral," *Fortune*, October 2, 2020, https://fortune.com/2020/10/02/deepfakes-putin-kim-jong-un-democracy-disinformation/; Phil Ehr (@PhilEhr), "@MattGaetz, you refuse to take Russia and election disinformation seriously. To get your attention, my campaign made a #DeepFake where you admit @BarackObama is cooler than you and @FoxNews sucks. If my campaign can do this, imagine what Putin is doing right now," Twitter, October 1, 2020, https://twitter.com/PhilEhr/status/1311667726742560769?s=19.

[105] National Intelligence Council, "Foreign Threats to the 2020 US Federal Elections," Office of the Director of National Intelligence, March 10, 2021, https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf; FBI Cyber Division, "Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations,"

Federal Bureau of Investigation, March 10, 2021, https://assets.documentcloud.org/documents/20514502/fbipin-3102021.pdf.

[106] Joan Solsman, "Samsung Deepfake AI Could Fabricate a Video of You From a Single Profile Pic," *CNET*, May 24, 2019, https://www.cnet.com/news/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/; Samsung Newsroom, "Samsung Electronics Launches AI Center in Russia," *Samsung*, May 29, 2018, https://news.samsung.com/global/samsung-electronics-launches-ai-center-in-russia.

[107] Note: Samsung researchers innovated by breaking up the process into stages. First, they trained a neural network on a large set of videos of people. Then they used GAN techniques to learn from a few images of a specific individual, to create increasingly realistic video of that individual. For details, see: Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models," arXiv preprint arXiv:1905.08233v1 (2019), https://arxiv.org/abs/1905.08233v1; Hecate He, "Samsung AI Makes the Mona Lisa 'Speak,'" *Synced*, May 23, 2019, https://medium.com/syncedreview/samsung-ai-makes-the-mona-lisa-speak-bea2b8362c38.

[108] James Vincent, "TikTok Tom Cruise Deepfake Creator: Public Shouldn't Worry About 'One-Click Fakes,'" *The Verge*, March 5, 2021, https://www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher.

[109] Note: Researchers have finetuned open-source models such as Speaker Verification to Multispeaker Text-to-Speech Synthesis (SV2TTS) system for voice-cloning, on precisely curated small training datasets of voice clips with relatively low compute requirements. For details, see: Corentin Jamine, "Real-Time Voice Cloning," *GitHub*, June 25, 2019, https://github.com/CorentinJ/Real-Time-Voice-Cloning; Tully and Foster, "Repurposing Neural Networks to Generate Synthetic Media for Information Operations."

[110] Tim Hwang, "Deepfakes: A Grounded Threat Assessment" (Center for Security and Emerging Technology, July 2020), https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/; Christiano Lima, "The Technology 202: As Senators Zero in on Deepfakes, Some Experts Fear Their Focus is Misplaced," *The Washington Post*, August 6, 2021, https://www.washingtonpost.com/politics/2021/08/06/technology-202-senators-zero-deepfakes-some-experts-fear-their-focus-is-misplaced/.

[111] Jesse Damiani, "The Curious Case Of Brad Pitt On Clubhouse," *Forbes*, February 16, 2021, https://www.forbes.com/sites/jessedamiani/2021/02/16/the-curious-case-of-brad-pitt-on-clubhouse/.

[112] Tully and Foster, “Repurposing Neural Networks to Generate Synthetic Media for Information Operations.”

[113] Sedova et al., “AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework.”

[114] Sensity Team, "How to Detect a Deepfake Online.”

[115] Robert Chesney and Danielle Citron, “Deepfakes and the New Disinformation War,” *Foreign Affairs*, January/February 2019, https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war; Sarah Cahlan, “How Misinformation Helped Spark an Attempted Coup in Gabon,” *The Washington Post*, February 13, 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/; Nic Ker, “Is the Political Aide Viral Sex Video Confession Real or a Deepfake?,” *Malay Mail*, June 12, 2019, https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422; Reuters Staff, “Fact Check: Donald Trump Concession Video Not a 'Confirmed Deepfake,'” Reuters, January 11, 2021, https://www.reuters.com/article/uk-factcheck-trump-consession-video-deep/fact-check-donald-trump-concession-video-not-a-confirmed-deepfake-idUSKBN29G2NL.

[116] Philip N. Howard, Samuel Woolley, and Ryan Calo, “Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration,” *Journal of Information Technology & Politics* (2018), https://doi.org/10.1080/19331681.2018.1448735; Stefano Cresci, “A Decade of Bot Detection,” *Communications of the ACM* 63, no. 10 (October 2020): 72-83, https://cacm.acm.org/magazines/2020/10/247598-a-decade-of-social-bot-detection/fulltext.

[117] Yue Yin, Hanzhou Wu, and Xinpeng Zhang, “Neural Visual Social Comment on Image-Text Content,” *IETE Technical Review*, March 1, 2020, https://doi.org/10.1080/02564602.2020.1730714.

[118] Dennis Assenmacher et al., “Demystifying Social Bots: On the Intelligence of Automated Social Media Actors,” *Social Media + Society*, September 1, 2020, https://doi.org/10.1177/2056305120939264.

[119] Carlos Freitas, Fabricio Benevenuto, Saptarshi Ghosh, and Adriano Veloso, “Reverse Engineering Socialbot Infiltration Strategies in Twitter,” *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (August 2015), https://doi.org/10.1145/2808797.2809292.

[120] “Datasets,” Botometer Bot Repository, https://botometer.osome.iu.edu/bot-repository/datasets.html.

[121] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications* 60 (2016): 19-31, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2015.11.016.

[122] Kai-Cheng Yang et al., "Arming the Public with Artificial Intelligence to Counter Social Bots," *Human Behavior and Emerging Technologies* 1, no. 1 (February 6, 2019), https://doi.org/10.1002/hbe2.115.

[123] Yang et al., "Arming the Public with Artificial Intelligence to Counter Social Bots."

[124] Will Knight, "This AI Can Generate Convincing Text—and Anyone Can Use It," *WIRED*, March 29, 2021, https://www.wired.com/story/ai-generate-convincing-text-anyone-use-it/.

[125] Assenmacher et al., "Demystifying Social Bots."

[126] Adam Rawnsley, "Right-Wing Media Outlets Duped by a Middle East Propaganda Campaign," *The Daily Beast*, July 7, 2020, https://www.thedailybeast.com/right-wing-media-outlets-duped-by-a-middle-east-propaganda-campaign.

[127] Frederik J. Zuiderveen Borgesius et al., "Should We Worry About Filter Bubbles?," *Internet Policy Review: Journal on Internet Regulation* 5, no. 1 (March 31, 2016), https://policyreview.info/node/401/pdf%0Ahttps:/pure.uva.nl/ws/files/21231009/Should_we_worry_about_filter_bubbles.pdf.

[128] Eli Pariser, *The Filter Bubble: What the Internet is Hiding from You* (New York, NY: Penguin Books, 2011).

[129] Aleksandar Ilic and Maja Kabijo, "Recommending Items to More than a Billion People," *Facebook Engineering*, June 2, 2015, https://engineering.fb.com/2015/06/02/core-data/recommending-items-to-more-than-a-billion-people/.

[130] Akos Lada, Meihong Wang, and Tak Yan, "How Machine Learning Powers Facebook's News Feed Ranking Algorithm," *Facebook Engineering*, January 26, 2021, https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/.

[131] Nicolas Koumchatzky and Anton Andryeyev, "Using Deep Learning at Scale in Twitter's Timelines," *Twitter Engineering*, May 9, 2017, https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines; Paul Covington, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings*

of the 10th ACM Conference on Recommender Systems (September 7, 2016), https://dl.acm.org/doi/10.1145/2959100.2959190.

[132] Guillaume Chaslot, "The Toxic Potential of YouTube's Feedback Loop," *WIRED*, July 13, 2019, https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/.

[133] Chaslot, "The Toxic Potential of YouTube's Feedback Loop."

[134] Manoel Horta Ribeiro et al., "Auditing Radicalization Pathways on YouTube," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (January 27, 2020): 131-141, https://doi.org/10.1145/3351095.3372879.

[135] DeepMind (@DeepMind), "Feedback loops in recommendation systems can give rise to 'echo chambers' and 'filter bubbles' which can narrow a user's content exposure, and ultimately shift their world view," Twitter, March 1, 2019, https://twitter.com/DeepMind/status/1101514121563041792; Ray Jiang et al., " Degenerate Feedback Loops in Recommender Systems," arXiv preprint arXiv:1902.10730 (2019), https://arxiv.org/pdf/1902.10730.pdf.

[136] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Pola, "Shilling Attacks Against Recommender Systems: A Comprehensive Survey," *Artificial Intelligence Review* 42 (December 2014), https://doi.org/10.1007/s10462-012-9364-9.

[137] Michael Golebiewski and danah boyd, "Data Voids: Where Missing Data Can Easily Be Exploited," *Data&Society*, October 29, 2019, https://datasociety.net/library/data-voids/.

[138] Golebiewski and boyd, "Data Voids: Where Missing Data Can Easily Be Exploited."

[139] Andrew Lohn, "Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity" (Center for Security and Emerging Technology, December 2020), https://cset.georgetown.edu/publication/hacking-ai/.

[140] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572 (2015), https://arxiv.org/abs/1412.6572.

[141] Rami Cohen, Dietmar Jannach, Oren Sar Shalom, and Amihood Amir, "A Black-Box Attack Model for Visually-Aware Recommender Systems," arXiv preprint arXiv:2011.02701 (2020), https://arxiv.org/pdf/2011.02701.pdf.

[142] Adam Mosseri, "Facebook Recently Announced a Major Update to News Feed; Here's What's Changing," *Facebook*, April 8, 2018, https://about.fb.com/news/2018/04/inside-feed-meaningful-interactions/; Akos Lada, Meihong Wang, and Tak Yan, "How Machine Learning Powers Facebook's

News Feed Ranking Algorithm," *Facebook Engineering*, January 26, 2021, https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/.

[143] Nina Jankowicz and Cindy Otis, "Facebook Groups Are Destroying America," *WIRED*, June 17, 2020, https://www.wired.com/story/facebook-groups-are-destroying-america/; Kate Linebaugh and Ryan Knutson, "The Facebook Files, Part 4: The Outrage Algorithm," *The Wall Street Journal*, September 18, 2021, https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/E619FBB7-43B0-485B-877F-18A98FFA773F.

[144] Tom Alison, "Changes to Keep Facebook Groups Safe," *Facebook Newsroom* March 17, 2021, https://about.fb.com/news/2021/03/changes-to-keep-facebook-groups-safe/.

[145] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[146] Ari Sen and Brandy Zadrozny, "QAnon Groups Have Millions of Members on Facebook, Documents Show," *NBC News*, August 10, 2020, https://www.nbcnews.com/tech/tech-news/qanon-groups-have-millions-members-facebook-documents-show-n1236317.

[147] The YouTube Team, "Continuing Our Work to Improve Recommendations on YouTube," *YouTube Official Blog*, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve/; Julia Alexander, "YouTube Claims Its Crackdown on Borderline Content is Actually Working," *The Verge*, December 3, 2019, https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources; Alison, "Changes to Keep Facebook Groups Safe"; The YouTube Team, "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation," *YouTube Official Blog*, December 3, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce.

[148] Representatives of Facebook, Twitter, and YouTube, "Algorithms and Amplification: How Social Media Platforms' Design Choices Shape Our Discourse and Our Minds," United States Senate Committee on the Judiciary Hearing, April 27, 2021, https://www.judiciary.senate.gov/meetings/algorithms-and-amplification-how-social-media-platforms-design-choices-shape-our-discourse-and-our-minds.

[149] Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," *Facebook*, May 5, 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[150] Melanie Smith, "Interpreting Social Qs: Implications of the Evolution of QAnon" (Graphika, August 2020), https://public-assets.graphika.com/reports/graphika_report_interpreting_social_qs.pdf.

[151] Renee DiResta, "Anti-Vaxxers Think This Is Their Moment," *The Atlantic*, December 20, 2020, https://www.theatlantic.com/ideas/archive/2020/12/campaign-against-vaccines-already-under-way/617443/; Elizabeth Dwoskin, "Massive Facebook Study on Users' Doubt in Vaccines Finds a Small Group Appears to Play a Big Role in Pushing the Skepticism," *The Washington Post*, March 14, 2021, https://www.washingtonpost.com/technology/2021/03/14/facebook-vaccine-hesistancy-qanon/.

[152] Kris McGuffie and Alex Newhouse, "The Radicalization Risks Posed by GPT-3 and Advanced Neural Language Models," Middlebury Institute of International Studies at Monterey, September 9, 2020, https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf; Buchanan et al., "Truth, Lies, and Automation."

[153] Edward Tian, "The QAnon Timeline: Four Years, 5,000 Drops and Countless Failed Prophecies," *Bellingcat*, January 29, 2021, https://www.bellingcat.com/news/americas/2021/01/29/the-qanon-timeline/.

[154] IBM Cloud Education, "Conversational AI," *IBM*, August 31, 2020, https://www.ibm.com/cloud/learn/conversational-ai.

[155] Nuacem AI, "An Expert System: Conversational AI vs Chatbots," *Chatbots Life*, March 19, 2021, https://chatbotslife.com/an-expert-system-conversational-ai-vs-chatbots-6b7c17c99258.

[156] Daniel Adiwardana and Thang Luong, "Towards a Conversational Agent that Can Chat About...Anything," *Google AI Blog*, January 28, 2020, https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html.

[157] Daniel Adiwardana et al., "Towards a Human-like Open-Domain Chatbot," arXiv preprint arXiv:2001.09977 (2020), https://arxiv.org/pdf/2001.09977.pdf.

[158] Yizhe Zhang et al., "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation" (Microsoft Research, November 2019), https://www.microsoft.com/en-us/research/publication/dialogpt-large-scale-generative-pre-training-for-conversational-response-generation/.

[159] Thomas Wolf, "How to Build a State-of-the-Art Conversational AI with Transfer Learning," *Hugging Face*, May 9, 2019, https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313; Tom Simonite, "To See the Future of Disinformation, You Build Robo-Trolls," *WIRED*, November 19,

2019, https://www.wired.com/story/to-see-the-future-of-disinformation-you-build-robo-trolls/.

[160] Stephen Roller, Emily Dinan, and Jason Weston, "A State-of-the-Art Open Source Chatbot," *Facebook AI Research*, April 29, 2020, https://ai.facebook.com/blog/state-of-the-art-open-source-chatbot/; Eric Michael Smith et al., "Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills," arXiv preprint arXiv:2004.08449 (2020), https://arxiv.org/pdf/2004.08449.pdf.

[161] Siqi Bao et al., "PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation," arXiv preprint arXiv:2109.09519 (2021), https://arxiv.org/abs/2109.09519.

[162] Hal Berghel and Daniel Berleant, "The Online Trolling Ecosystem," *Computer*, August 2018, http://www.berghel.net/col-edit/aftershock/aug-18/aftershock_8-18.pdf.

[163] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[164] Samantha Cole, "Microsoft Used Machine Learning to Make a Bot That Comments on News Articles For Some Reason," *VICE*, October 4, 2019, https://www.vice.com/en/article/d3a4mk/microsoft-used-machine-learning-to-make-a-bot-that-comments-on-news-articles-for-some-reason.

[165] Jasper Linmans, Bob van de Velde, and Evangelos Kanoulas, "Improved and Robust Controversy Detection in General Web Pages Using Semantic Approaches under Large Scale Conditions," arXiv preprint arXiv:1812.00382 (2018), https://arxiv.org/pdf/1812.00382.pdf; Dilek Kucuk and Fazli Can, "Stance Detection: A Survey," *ACM Computing Surveys* 53, no. 1 (February 2020), https://dl.acm.org/doi/pdf/10.1145/3369026.

[166] IBM Watson Discovery, "Find Critical Answers and Insights from Your Business Data using AI-Powered Enterprise Search Technology," *IBM*, https://www.ibm.com/cloud/watson-discovery.

[167] Nisarg Rajpura, Alizain Tejani, Sujal Upadhyay, and Rovina D'Britto, "Suicidal Tendencies and Ideation Prediction using Reddit," *International Research Journal of Engineering and Technology* 8, no. 5 (May 5, 2021), https://www.irjet.net/archives/V8/i5/IRJET-V8I5788.pdf.

[168] Nathan P. Kalmoe, "Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence," *Political Communication*, 31, no. 4 (October 16, 2014): 545–63, https://doi.org/10.1080/10584609.2013.852642.

[169] McGuffie and Newhouse, "The Radicalization Risks Posed by GPT-3 and Advanced Neural Language Models"; Buchanan et al., "Truth, Lies, and Automation."

[170] Facebook for Developers, "Bots for Workplace: Building Bots for Workplace in Groups and Chat," *Facebook*, https://developers.facebook.com/docs/workplace/integrations/custom-integrations/bots/; Open Data Science, "What Does Facebook's Blenderbot 2.0 Mean for the Future of AI?" *ODSC Medium Blog*, September 17, 2021, https://medium.com/@ODSC/what-does-facebooks-blenderbot-2-0-mean-for-the-future-of-ai-13b195dbd6ab.

[171] Cade Metz, "Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected,'" *The New York Times*, June 16, 2020, https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html.

[172] Sophie Webster, "Replika AI Creates Platform That Trains Chatbots, Similar to Black Mirror's 'Be Right Back' Episode," *Tech Times*, July 8, 2021, https://www.techtimes.com/articles/262609/20210708/replika-ai-creates-platform-trains-chatbots-similar-black-mirrors-right.htm.

[173] Metz, "Riding Out Quarantine With a Chatbot Friend."

[174] Tamaghna Basu, "How I Clone Myself Using AI," *BlackHat Security Conference*, August 2020, https://i.blackhat.com/USA-20/Thursday/us-20-Basu-How-I-Created-My-Clone-Using-AI-Next-Gen-Social-Engineering.pdf.

[175] James Vincent, "'Deepfake' that Supposedly Fooled European Politicians Was Just a Look-Alike, Say Pranksters: Fear of Deepfakes Seems to Have Outpaced the Technology Itself," *The Verge*, April 30, 2021, https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus.

[176] Samantha Cole, "This Open-Source Program Deepfakes You During Zoom Meetings, in Real Time," *VICE*, April 16, 2020, https://www.vice.com/en/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time.

[177] Aliaksandr Siarohin et al., "First Order Motion Model for Image Animation," arXiv preprint arXiv:2003.00196 (2020), https://arxiv.org/pdf/2003.00196.pdf.

[178] Ali Aliev, "Avarify Python: Avatars for Zoom, Skype and Other Video-Conferencing Apps," *GitHub*, July 24, 2020, https://github.com/alievk/avatarify-python#readme.

[179] Hannah Arendt, "Hannah Arendt: From an Interview," *The New York Review*, October 26, 1978, https://www.nybooks.com/articles/1978/10/26/hannah-arendt-from-an-interview/.

[180] Zeng et al., "PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation."

[181] Jimit Bagadiya, "367 Social Media Statistics You Must Know In 2021," *SocialPilot*, https://www.socialpilot.co/blog/social-media-statistics.

[182] Smith, "Interpreting Social Qs: Implications of the Evolution of QAnon."

[183] Bradshaw, Bailey, and Howard, "Disinformation: 2020 Global Inventory of Organized Social Media Manipulation."

[184] Gleicher et al., "The State of Influence Operations 2017-2020"; Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior," *Facebook*, October 8, 2020, https://about.fb.com/news/2020/10/removing-coordinated-inauthentic-behavior-september-report/; "October 2020 Coordinated Inauthentic Behavior Report," *Facebook*, November 4, 2020, https://about.fb.com/news/2020/11/october-2020-cib-report/; "December 2020 Coordinated Inauthentic Behavior Report," *Facebook*, January 12, 2021, https://about.fb.com/news/2021/01/december-2020-coordinated-inauthentic-behavior-report/.

[185] Graphika Team, "Facebook's Roger Stone Takedown: Facebook Removes Inauthentic Network Attributed to Political Operative" (Graphika, July 2020), https://graphika.com/reports/facebooks-roger-stone-takedown/; Tal Axelrod, "Twitter Suspends 70 Pro-Bloomberg 'Spam' Accounts," *The Hill*, February 21, 2020, https://thehill.com/policy/technology/484168-twitter-suspends-70-pro-bloomberg-accounts-citing-platform-manipulation; Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory, "The Long Fuse: Misinformation and the 2020 Election" (Election Integrity Partnership, 2021), https://purl.stanford.edu/tr171zs0069.

[186] Gleicher, "Removing Coordinated Inauthentic Behavior," *Facebook*, October 8, 2020; "October 2020 Coordinated Inauthentic Behavior Report," *Facebook*; Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior," *Facebook*, July 8, 2020, https://about.fb.com/news/2020/07/removing-political-coordinated-inauthentic-behavior/; Graphika Team, "Facebook's Roger Stone Takedown: Facebook Removes Inauthentic Network Attributed to Political Operative."

[187] Sara Fischer and Ashley Gold, "All the Platforms That Have Banned or Restricted Trump So Far," *Axios*, January 11, 2021, https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html.

[188] Isaac Stanley-Becker, "Facebook Bans Rally Forge, Marketing Firm Working on Behalf of Turning Point USA Affiliate," *The Washington Post*, October 8, 2020, https://www.washingtonpost.com/technology/2020/10/08/facebook-bans-media-consultancy-running-troll-farm-pro-trump-youth-group/.

[189] Jack Dorsey (@jack), "We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought. Why? A few reasons. . ." *Twitter*, October 30, 2019, https://twitter.com/jack/status/1189634360472829952?s=20.

[190] Note: Top AI research institutions have hundreds of public source code repositories on GitHub. DeepMind has 98: https://github.com/deepmind; OpenAI has 116 including GPT-2: https://github.com/openai; and Google Research has 123: https://github.com/google-research.

[191] Ng Wai Foong, "Beginner's Guide to Retrain GPT-2 (117M) to Generate Custom Text Content," *Medium*, May 12, 2019, https://medium.com/ai-innovation/beginners-guide-to-retrain-gpt-2-117m-to-generate-custom-text-content-8bb5363d8b7f; Code Mental, "Creating an AI Deepfake Version of Me With Voice Using Wav2Lip and Google Wavenet," *YouTube*, December 4, 2020, https://www.youtube.com/watch?v=DRXFcT48JqY.

[192] "This X Does Not Exist," https://thisxdoesnotexist.com/.

[193] Corentin Jamine, "Real-Time-Voice-Cloning: Clone a Voice in 5 seconds to Generate Arbitrary Speech in Real-Time," *GitHub*, https://github.com/CorentinJ/Real-Time-Voice-Cloning#readme.

[194] Catalin Cimpanu, "Malware Gangs Love Open Source Offensive Hacking Tools," *ZDNet*, October 12, 2020, https://www.zdnet.com/article/malware-gangs-love-open-source-offensive-hacking-tools/.

[195] Joan E. Solsman, "YouTube's AI is the Puppet Master Over Most of What You Watch," *CNET*, January 10, 2018, https://www.cnet.com/news/youtube-ces-2018-neal-mohan/; Aaron Smith, Skye Toor, and Patrick Van Kessel, "Many Turn to YouTube for Children's Content, News, How-To Lessons," *Pew Research Center*, November 7, 2018, https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/#an-analysis-of-random-walks-through-the-youtube-recommendation-engine; Zeynep Tufekci, "YouTube, the Great Radicalizer," *The New York Times*, March 10, 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html; Galen Stocking et al., "Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side," *Pew Research Center*, September 28, 2020, https://www.journalism.org/wp-content/uploads/sites/8/2020/09/Many-Americans-Get-News-on-YouTube-

Where-News-Organizations-and-Independent-Producers-Thrive-Side-by-Side.pdf.

[196] The YouTube Team, "Continuing Our Work to Improve Recommendations on YouTube," *YouTube Official Blog*, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve; Spandana Singh, "Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences" (Open Technology Institute, New America, March 25, 2020), https://www.newamerica.org/oti/reports/why-am-i-seeing-this/case-study-youtube; Julia Alexander, "YouTube Claims Its Crackdown on Borderline Content is Actually Working," *The Verge*, December 3, 2019, https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources; Jonas Kaiser and Adrian Rauchfleisch, "How YouTube Helps Form Homogenous Online Communities," *Brookings*, December 23, 2020, https://www.brookings.edu/techstream/how-youtube-helps-form-homogeneous-online-communities/; Jonas Kaiser and Adrian Rauchfleisch, "Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube's Channel Recommendations in the United States and Germany," *Social Media + Society* (2020), https://doi.org/10.1177/2056305120969914.

[196] Kaiser and Rauchfleisch, "How YouTube Helps Form Homogenous Online Communities."

[197] Sedova et al., "AI and the Future of Disinformation Campaigns: Part 1, The RICHDATA Framework."

[198] "Remarks by National Security Advisor Jake Sullivan at the National Security Commission on Artificial Intelligence Global Emerging Technology Summit," The White House, July 13, 2021, https://www.whitehouse.gov/nsc/briefing-room/2021/07/13/remarks-by-national-security-advisor-jake-sullivan-at-the-national-security-commission-on-artificial-intelligence-global-emerging-technology-summit/.

[199] Müge Fazlioglu, "Privacy Bills in the 117th Congress," *International Association of Privacy Professionals*, August 24, 2021, https://iapp.org/news/a/privacy-bills-in-the-117th-congress/.

[200] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," arXiv preprint arXiv:1906.04043 (2019), https://arxiv.org/abs/1906.04043; Identifying Outputs of Generative Adversarial Networks Act, S.2904, 116th Cong. (2020), https://www.congress.gov/bill/116th-congress/senate-bill/2904; Matt Turek, "Semantic Forensics (SemaFor)," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/semantic-forensics; Laurie A. Harris, "Deep Fakes and National Security" (Congressional Research Service, June 8, 2021), https://crsreports.congress.gov/product/pdf/IF/IF11333.

[201] "Content Authenticity Initiative," Content Authenticity, 2019, https://contentauthenticity.org/; Tom Burt and Eric Horvitz, "New Steps to Combat Disinformation," *Microsoft Corporation*, September 1, 2020, https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/; Jigsaw and Google Research, "Project Assembler," Project Assembler, https://projectassembler.org/; "The Coalition for Content Provenance and Authenticity (C2PA)," February 2021, https://c2pa.org/; Microsoft News Center, "Technology and Media Entities Join Forces to Create Standards Group Aimed at Building Trust in Online Content," *Microsoft*, February 22, 2021, https://news.microsoft.com/2021/02/22/technology-and-media-entities-join-forces-to-create-standards-group-aimed-at-building-trust-in-online-content/.

[202] Andrew Imbrie, Ryan Fedasiuk, Catherine Aiken, Tarun Chhabra, and Husanjot Chahal, "Agile Alliances: How the United States and Its Allies Can Deliver a Democratic Way of AI" (Center for Security and Emerging Technology, February 2020), https://cset.georgetown.edu/research/agile-alliances/; Andrew Imbrie and Ryan Fedasiuk, "An Alliance-Centered Approach to AI" (Center for Security and Emerging Technology, September 2020), https://cset.georgetown.edu/research/an-alliance-centered-approach-to-ai/.

[203] Note: Select rapid response solutions include but are not limited to: European Council's Rapid Alert System (RAS), European Hybrid Warfare Center of Excellence, NATO's Strategic Communications and Cyber Centers of Excellence, EU's East StratCom Taskforce among others.

[204] "Information Sharing and Analysis Organizations (ISAOS)," Cybersecurity and Infrastructure Security Agency, https://www.cisa.gov/information-sharing-and-analysis-organizations-isaos.

[205] Fischer and Gold, "All the Platforms That Have Banned or Restricted Trump So Far."

[206] A recently announced DARPA project, Influence Campaign Awareness and Sensemaking (INCAS), may fill this gap to increase awareness of global influence campaigns in information environments outside the U.S., but is prohibited from monitoring the U.S. domestic context. For details, see: Brian Kettler, "Influence Campaign Awareness and Sensemaking," Defense Advanced Research Projects Agency, October 26, 2020, https://www.darpa.mil/program/influence-campaign-awareness-and-sensemaking.

[207] Nathaniel Gleicher, "Recommended Principles for Regulation or Legislation to Combat Influence Operations," *Facebook*, October 8, 2020, https://about.fb.com/news/2020/10/recommended-principles-for-regulation-or-legislation-to-combat-influence-operations/; Amy O'Hara and Jodi Nelson, "Evaluation of the Social Science One - Social Science Research Council - Facebook Partnership," Hewlett Foundation, December 2019,

https://hewlett.org/wp-content/uploads/2020/02/Facebook-Partnership-Final-Evaluation-Report.pdf.

[208] Note: Some examples include cooperation on cybersecurity of critical infrastructure and counterterrorism, such as Information Sharing and Analysis Organizations (ISAOs) and Global Internet Forum to Counter Terrorism (GIFCT), as well as Election Infrastructure Information Sharing and Analysis Center (EI-ISAC) and the privately funded Election Integrity Partnership for cooperation during elections. For details, see: "Information Sharing and Analysis Organizations (ISAOS)," Cybersecurity and Infrastructure Security Agency; "ISAO 100-1: Introduction to Information Sharing and Analysis Organizations (ISAOs)," ISAO Standards Organization, October 14, 2016, https://www.isao.org/wp-content/uploads/2016/10/ISAO-100-1-Introduction-to-ISAO-v1-01_Final.pdf; Global Internet Forum to Counter Terrorism, https://gifct.org/about/; Antigone Davis and Guy Rosen, "Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer," *Facebook*, August 1, 2019, https://about.fb.com/news/2019/08/open-source-photo-video-matching/; Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory, "The Long Fuse: Misinformation and the 2020 Election."

[209] Jack Clark, "Import AI 243: Training AI with Fractals, RL-Trained Walking Robots, and the European AI Fund Makes Grants to 16 Organizations," *ImportAI Newsletter*, April 5, 2021, https://jack-clark.net/2021/04/05/import-ai-243-training-ai-with-fractals-rl-trained-walking-robots-and-the-european-ai-fund-makes-grants-to-16-organizations/.

[210] Note: The Montreal Declaration on Responsible AI is an attempt by Canadian academia and AI industry to set basic norms for an ethically responsible development of AI. More than 1,200 researchers signed the Future of Life's Asilomar AI Principles, stating that "designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications." For details, see: "The Montreal Declaration for Responsible AI," University of Montreal, https://www.montrealdeclaration-responsibleai.com/the-declaration; "Asilomar AI Principles," Future of Life Institute, https://futureoflife.org/ai-principles/; "Better Language Models and Their Implications," *OpenAI Blog*, February 14, 2019, https://openai.com/blog/better-language-models/; Rowan Zellers et al., "Defending Against Neural Fake News," arXiv preprint arXiv:1905.12616 (2020), https://arxiv.org/pdf/1905.12616.pdf.

[211] "Call for Papers," Neural Information Processing Systems (NeurIPS) Conference, 2020, https://nips.cc/Conferences/2020/CallForPapers.

[212] Clement Delangue, "Ethical Analysis of the Open-Sourcing of a State-Of-the-Art Conversational AI," *Hugging Face*, May 9, 2019,

https://medium.com/huggingface/ethical-analysis-of-the-open-sourcing-of-a-state-of-the-art-conversational-ai-852113c324b2.

[213] Connor Leahy, "Why Release a Large Language Model?"

[214] Toby Shevlane and Allan Dafoe, "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" arXiv preprint arXiv:2001.00463 (2019), https://arxiv.org/abs/2001.00463.

[215] "Asilomar AI Principles," Future of Life Institute, https://futureoflife.org/ai-principles/.

[216] Jay Rosen, "The Big National News Providers Need Threat Modeling Teams," PressThink, September 14, 2020, https://pressthink.org/2020/09/the-national-news-providers-need-threat-modeling-teams/; Jay Rosen, "What newsrooms can learn from threat modeling at Facebook," *The Verge*, September 14, 2020, https://www.theverge.com/21435639/threat-modeling-facebook-alex-stamos-newsroom-security.

[217] Jigsaw and Google Research, "Project Assembler."

[218] "Managing the COVID-19 Infodemic: Promoting Healthy Behaviours and Mitigating the Harm from Misinformation and Disinformation," World Health Organization, September 23, 2020, https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

[219] P.W. Singer and Michael McConnell, "Want to Stop the Next Crisis? Teaching Cyber Citizenship Must Become a National Priority," *TIME*, January 21, 2021, https://time.com/5932134/cyber-citizenship-national-priority/; "Learn to Discern (L2D) - Media Literacy Training," *IREX*, https://www.irex.org/project/learn-discern-l2d-media-literacy-training#component-id-783.

[220] "Tools That Fight Disinformation Online," Fighting Disinformation Project, RAND Corporation, https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html#q=&typeOfTool=Education%2Ftraining.

[221] Jeff Kosseff and Daphne Keller, "Why Outlawing Harmful Social Media Content Would Face an Uphill Legal Battle," *The Washington Post*, October 9, 2021, https://www.washingtonpost.com/outlook/2021/10/09/facebook-algorithm-first-amendment/.

[222] Eric Lander and Alondra Nelson, "Americans Need a Bill of Rights for an AI-Powered World: The White House Office of Science and Technology Policy is Developing Principles to Guard Against Powerful Technologies with Input From

the Public," *WIRED*, October 8, 2021, https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/.

[223] Claire Wardle, "Understanding Information Disorder" (First Draft, October 2019), https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf.