Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations

Josh A. Goldstein[1,3], Girish Sastry[*2], Micah Musser[*1],
Renée DiResta[3], Matthew Gentzel[2], and Katerina Sedova[1]

[1] *Georgetown University's Center for Security and Emerging Technology*
[2] *OpenAI*
[3] *Stanford Internet Observatory*

January 2023

Workshop Participants: Steven Adler, Shahar Avin, John Bansemer, Chris Bregler, Miles Brundage, Sam Gregory, Shelby Grossman, Ariel Herbert-Voss, Yacine Jernite, Claire Leibowicz, Connor Leahy, Herb Lin, Drew Lohn, Meg Mitchell, Amnon Morag, Alex Newhouse, Helen Ngo, Aviv Ovadya, Cooper Raterink, Yoel Roth, Bob Rotsted, Elizabeth Seger, and Raymond Serrato.

Acknowledgements: We thank participants in the October 2021 workshop that we convened for informing our understanding of various threats and mitigations. We also thank many workshop participants for providing feedback on a draft of this paper. For additional feedback on the paper, we thank Deepesh Chaudhari, Jeff Ding, Tyna Elondou, Shengli Hu, Daniel Kokotajlo, Gretchen Krueger, Pamela Mishkin, Ronald Robertson, Sarah Shoker, Samuel Wolrich, and Jenny Xiao. Josh Goldstein began working on the project as a postdoctoral fellow at Stanford, and continued work as a research fellow with Georgetown CSET's CyberAI Project. Matthew Gentzel completed his contributions while contracting for OpenAI, and is now at Longview Philanthropy. Katerina Sedova completed her contributions to this project while she was a research fellow with Georgetown CSET's CyberAI Project and before she entered U.S. government service. All errors remain our own.

Lead authors contributed equally.

Contents

Executive Summary

In recent years, artificial intelligence (AI) systems have significantly improved and their capabilities have expanded. In particular, AI systems called "generative models" have made great progress in automated content creation, such as images generated from text prompts. One area of particularly rapid development has been generative models that can produce original language, which may have benefits for diverse fields such as law and healthcare.

However, there are also possible negative applications of generative language models, or "language models" for short. For malicious actors looking to spread propaganda—information designed to shape perceptions to further an actor's interest—these language models bring the promise of automating the creation of convincing and misleading text for use in influence operations, rather than having to rely on human labor. For society, these developments bring a new set of concerns: the prospect of highly scalable—and perhaps even highly persuasive—campaigns by those seeking to covertly influence public opinion.

This report aims to assess: how might language models change influence operations, and what steps can be taken to mitigate these threats? This task is inherently speculative, as both AI and influence operations are changing quickly.

Many ideas in the report were informed by a workshop convened by the authors in October 2021, which brought together 30 experts across AI, influence operations, and policy analysis to discuss the potential impact of language models on influence operations. The resulting report does not represent the consensus of workshop participants, and mistakes are our own.

We hope this report is useful to disinformation researchers who are interested in the impact of emerging technologies, AI developers setting their policies and investments, and policymakers preparing for social challenges at the intersection of technology and society.

Potential Applications of Language Models to Influence Operations

We analyzed the potential impact of generative language models on three well-known dimensions of influence operations—the **actors** waging the campaigns, the deceptive **behaviors** leveraged as tactics, and the **content** itself—and conclude that language models could significantly affect how influence operations are waged in the future. These changes are summarized in Table 1.

The potential of language models to rival human-written content at low cost suggests that these models—like any powerful technology—may provide distinct advantages to propagandists who choose to use them. These advantages could expand access to a greater number of actors, enable new tactics of influence, and make a campaign's messaging far more tailored and potentially effective.

Progress in Influence Operations and Critical Unknowns

Technical progress in language models is unlikely to halt, so any attempt to understand how language models will affect future influence operations needs to take expected progress into account. Language models are likely to become more **usable** (making it easier to apply models to a task), **reliable** (reducing the chance that models produce outputs with obvious errors), and **efficient** (increasing the cost-effectiveness of applying a language model for influence operations).

Dimension[1]	Potential Change Due to Generative AI Text	Explanation of Change
Actors	Larger number and more diverse group of propagandists emerge.	As generative models drive down the cost of generating propaganda, more actors may find it attractive to wage influence operations.
	Outsourced firms become more important.	Propagandists-for-hire that automate the production of text may gain new competitive advantages.
Behavior	Automating content production increases scale of campaigns.	Propaganda campaigns will become easier to scale when text generation is automated.
	Existing behaviors become more efficient.	Expensive tactics like cross-platform testing may become cheaper with language models.
	Novel tactics emerge.	Language models may enable dynamic, personalized, and real-time content generation like one-on-one chatbots.
Content	Messages are more credible and persuasive.	Generative models may improve messaging compared to text written by propagandists who lack linguistic or cultural knowledge of their target.
	Propaganda is less discoverable.	Existing campaigns are frequently discovered due to their use of copy-and-pasted text (copypasta), but language models will allow the production of linguistically distinct messaging.

Table 1: How Language Models May Affect the ABCs of Influence Operations

These factors lead us to make a high confidence judgment that language models will be useful for influence operations in the future. The exact nature of their application, however, is unclear.

There are several critical unknowns that will impact how, and the extent to which, language models will be adopted for influence operations. These unknowns include:

- **Which new capabilities for influence will emerge as a side-effect of well-intentioned research?** The conventional research process—which targets more general language tasks—has resulted in systems that could be applied to influence operations. New capabilities, like producing longform persuasive arguments, could emerge in the future. These emergent capabilities are hard to anticipate with generative models, but could determine the specific tasks propagandists will use language models to perform.
- **Will it be more effective to engineer specific language models for influence operations, rather than apply generic ones?** While most current models are built for generic tasks or tasks of scientific or commercial value, propagandists could build or adapt models to be directly useful for tasks like persuasion and social engineering. For example, a propagandist may be able to adapt a smaller, less capable model in a process known as fine-tuning. This is likely cheaper than building a larger, more general model, though it is uncertain how much cheaper this would be. Furthermore, fine-tuning a state-of-the-art model could make novel capabilities for influence easier for propagandists to obtain.

[1] Dimension categories drawn from Camille François's "Disinformation ABC" framework.

- **Will actors make significant investments in language models over time?** If many actors invest in, and create, large language models, it will increase the likelihood of propagandists gaining access to language models (legitimately or via theft). Propagandists themselves could invest in creating or fine-tuning language models, incorporating bespoke data—such as user engagement data—that optimizes for their goals.
- **Will governments or specific industries create norms against using models for propaganda purposes?** Just as norms around use constrain the misuse of other technologies, they may constrain the application of language models to influence operations. A coalition of states who agree not to use language models for propaganda purposes could impose costs on those that fail to abide. On a substate level, research communities and specific industries could set norms of their own.
- **When will easy-to-use tools to generate text become publicly available?** Language models still require operational know-how and infrastructure to use skillfully. Easy-to-use tools that produce tweet- or paragraph-length text could lead existing propagandists who lack machine learning know-how to rely on language models.

Because these are critical possibilities that can change how language models may impact influence operations, additional research to reduce uncertainty is highly valuable.

What Can Be Done to Mitigate the Potential Threat?

Building on the workshop we convened in October 2021, and surveying much of the existing literature, we attempt to provide a kill chain framework for, and a survey of, the types of different possible mitigation strategies. Our aim is not to endorse specific mitigations, but to show how mitigations could target different stages of the influence operation pipeline.

What Propagandists Require	**Stage of Intervention**	**Illustrative Mitigations**
1. Language Models Capable of Producing Realistic Text	Model Design and Construction	AI Developers Build Models That Are More Fact-Sensitive
		Developers Spread Radioactive Data to Make Generative Models Detectable
		Governments Impose Restrictions on Data Collection
		Governments Impose Access Controls on AI Hardware
2. Reliable Access to Such Models	Model Access	AI Providers Impose Stricter Usage Restrictions on Language Models
		AI Developers Develop New Norms Around Model Release
3. Infrastructure to Distribute the Generated Content	Content Dissemination	Platforms and AI Providers Coordinate to Identify AI Content
		Platforms Require "Proof of Personhood" to Post
		Entities That Rely on Public Input Take Steps to Reduce Their Exposure to Misleading AI Content
		Digital Provenance Standards Are Widely Adopted
4. Susceptible Target Audience	Belief Formation	Institutions Engage in Media Literacy Campaigns
		Developers Provide Consumer Focused AI Tools

Table 2: Summary of Example Mitigations

The table above demonstrates that there is no silver bullet that will singularly dismantle the threat of language models in influence operations. Some mitigations are likely to be socially infeasible, while others will require technical breakthroughs. Others may introduce unacceptable downside risks. Instead, to effectively mitigate the threat, a whole of society approach, marrying multiple mitigations, will likely be necessary.

Furthermore, effective management will require a cooperative approach among different institutions such as AI developers, social media companies, and government agencies. Many proposed mitigations will have a meaningful impact only if these institutions work together. It will be difficult for social media companies to know if a particular disinformation campaign uses language models unless they can work with AI developers to attribute that text to a model. The most radical mitigations—such as inserting content provenance standards into the protocols of the internet—would require extreme coordination, if they are desirable at all.

Perhaps most importantly, the mitigations we highlight require much more development, scrutiny, and research. Evaluating their effectiveness and robustness is worthy of serious analysis.

1 Introduction

1.1 Motivation

In recent years, as the capabilities of generative artificial intelligence (AI) systems—otherwise known as "generative models"—have improved, commentators have hypothesized about both the potential benefits and risks associated with these models. On the one hand, generative AI systems open up possibilities in fields as diverse as healthcare, law, education, and science.[2] For example, generative models are being used to design new proteins,[3] generate source code,[4] and inform patients.[5] Yet the rapid speed of technological progress has made it difficult to adequately prepare for, or even understand, the potential negative externalities of these models. Early research has suggested that bias in model generations could exacerbate inequalities, that models could displace human workers, and that, in the wrong hands, models could be intentionally misused to cause societal harm.[6]

Concurrently, the last decade has seen a rise in political influence operations—covert or deceptive efforts to influence the opinions of a target audience—online and on social media platforms specifically. Researchers and social media platforms have documented hundreds of domestic and foreign influence operations that are designed to mislead target audiences.[7] In the United States, the US intelligence community has publicly stated that foreign governments, including Russia and Iran, have waged influence operations targeting the 2016 and 2020 US presidential elections.[8]

In this paper, we focus on the overlap between these two trends. First, we ask: How can language models, a form of generative AI that can produce original text, impact the future of influence operations? While several studies have addressed specific applications, we provide frameworks for thinking through different types of changes and highlight critical unknowns that will affect the ultimate impact. By highlighting the technology's current limitations and critical unknowns, we attempt to avoid threat inflation or a sole focus on doomsday scenarios. After developing the threats, we ask: What are the possible mitigation strategies to address these various threats?

Our paper builds on a yearlong collaboration between OpenAI, the Stanford Internet Observatory (SIO), and Georgetown's Center for Security and Emerging Technology (CSET). In October 2021, we convened

[2] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," *arxiv:2108.07258 [cs.LG]*, August 2021, https://doi.org/10.48550/arxiv.2108.07258.

[3] Mohammed AlQuraishi, "Machine learning in protein structure prediction," *Current Opinion in Chemical Biology* 65 (December 2021): 1–8, ISSN: 1367-5931, https://doi.org/10.1016/J.CBPA.2021.04.005.

[4] "ML-Enhanced Code Completion Improves Developer Productivity," Google AI Blog, accessed July 28, 2022, https://ai.googleblog.com/2022/07/ml-enhanced-code-completion-improves.html.

[5] Maguire Herriman et al., "Asked and Answered: Building a Chatbot to Address Covid-19-Related Concerns," *NEJM Catalyst Innovations in Care Delivery*, June 18, 2020, https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0230.

[6] See for example Mark Chen et al., "Evaluating Large Language Models Trained on Code," *arxiv:2107.03374 [cs.LG]*, July 14, 2021, https://doi.org/10.48550/arxiv.2107.03374; Bommasani et al., "On the Opportunities and Risks of Foundation Models"; Sarah Kreps, R. Miles McCain, and Miles Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *Journal of Experimental Political Science* 9, no. 1 (November 2022): 104–117, ISSN: 2052-2630, https://doi.org/10.1017/XPS.2020.37; Ben Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation* (Center for Security and Emerging Technology, May 2021), https://doi.org/10.51593/2021CA003.

[7] For a list of influence operations removed from Facebook alone, see Nathaniel Gleicher et al., *Threat Report: The State of Influence Operations 2017-2020* (Meta, May 2021), https://about.fb.com/news/2021/05/influence-operations-threat-report/

[8] National Intelligence Council, *Intelligence Community Assessment: Foreign Threats to the 2020 US Federal Elections* (National Intelligence Council, March 10, 2021), https://int.nyt.com/data/documenttools/2021-intelligence-community-election-interference-assessment/abd0346ebdd93e1e/full.pdf.

a two-day workshop among 30 disinformation and machine learning experts in industry and academia to discuss the emerging threat as well as potential mitigations. This paper builds on the whitepaper that we circulated to workshop participants, the workshop itself, and subsequent months of research. We thank workshop participants for helping to clarify potential vectors of abuse and possible mitigations, and note that our report does not necessarily reflect the views of the participants.

1.2 Threats and Mitigations

How can language models affect the future of influence operations?

To address this question, we build on the ABC model — Actors, Behaviors, and Content — from the disinformation literature.[9] Language models can affect *which* actors wage influence operations, *how* they do so, and *what* content they produce.

- **Actors:** Language models drive down the cost of generating propaganda—the deliberate attempt to shape perceptions and direct behavior to further an actor's interest[10]—so more actors may find it attractive to wage these campaigns.[11] Likewise, propagandists-for-hire that automate production of text may gain new competitive advantages.
- **Behavior:** Recent AI models can generate synthetic text that is highly scalable, and often highly persuasive.[12] Influence operations with language models will become easier to scale, and more expensive tactics (e.g., generating personalized content) may become cheaper. Moreover, language models could enable new tactics to emerge—like real-time content generation in one-on-one chatbots.
- **Content:** Language models may create more impactful messaging compared to propagandists who lack linguistic or cultural knowledge of their target. They may also make influence operations less discoverable, since they create new content with each generation.

When considering these predicted changes, it is also important to remember that AI development is progressing rapidly. We highlight critical unknowns that will impact the future of influence operations, including how models will improve, whether new capabilities will emerge as a product of scale, whether actors invest in AI for influence operations, and whether norms emerge that constrain different actors from automating their influence campaigns.

What mitigations could reduce the impact of AI-enabled influence operations?

After laying out potential threats, we also consider the range of possible mitigation strategies to influence operations with language models. We develop a framework that categorizes mitigations based on a kill

[9] Camille François, *Actors, Behaviors, Content: A Disinformation ABC Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, September 2019), https://science.house.gov/download/francois-addendum.

[10] Garth Jowett and Victoria O'Donnell, *Propaganda & Persuasion*, 6th ed. (SAGE Publications, 2014), ISBN: 1483323528; Philip M. Taylor, *Munitions of the mind: a history of propaganda from the ancient world to the present era* (Manchester University Press, 2003), ISBN: 978-1-84779-092-7.

[11] We include a rough cost-effectiveness calculation in Section 4.1.3; see also Micah Musser, "A Cost Analysis of Generative Language Models and Influence Operations," *(Working Paper)*.

[12] Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation*; Josh A. Goldstein et al., "Can AI write persuasive propaganda?," *(Working Paper)*.

chain framework. To effectively wage an influence operation with a language model, propagandists would require (1) that a model is built (by themselves or others), (2) that they have access to the model, (3) that they have the means of disseminating content they produce, and (4) that the information spread impacts the target. Each of these steps—model design and construction, model access, content dissemination, and belief formation—represents a possible stage for intervention.

1.3 Scope and Limitations

This paper focuses on a particular application of AI (language models) to influence operations, but it does not focus on other AI models, other forms of information control, or specific actors. As described above, generative models include models that can create a range of output. The idea of AI-generated "deepfaked" images or video has been in the public consciousness for several years now.[13] Recently, for example, a low-quality deepfake video of Ukrainian President Volodymyr Zelensky purportedly telling Ukrainian soldiers to lay down their arms and surrender circulated on social media.[14] Higher-quality deepfake videos have also gained traction in the past.[15] We focus on generative text, rather than videos, images, or multimodal models for three reasons: first, because text is relatively underexplored (compared to images and videos) in the disinformation literature, second, because text seems particularly difficult to distinguish as AI-generated, and third, because access to these capabilities is diffusing quickly.[16] While multimodal models are also new and relatively underexplored, they are not our primary focus.

Our focus on how language models can be used for influence operations scopes our study more narrowly than information control writ large. State and non-state actors engage in a variety of information control behaviors, ranging from censorship to manipulating search algorithms. One recent framework categorizes different forms of digital repression, and notes that these techniques are as distinct as "online disinformation campaigns, digital social credit schemes, private online harassment campaigns by lone individuals, and regime violence against online political actors."[17] While we take digital repression seriously, a fuller examination of categories of digital repression other than covert propaganda campaigns—and how those categories are affected by AI—falls outside our scope.

Our scope is relevant to a variety of state, substate, and private actors; we do not focus on any one actor specifically. Although the intentions and capabilities of specific actors is relevant to assess the likelihood

[13] Claire Wardle, "This Video May Not Be Real," *New York Times*, August 19, 2019, https://www.nytimes.com/2019/08/14/opinion/deepfakes-adele-disinformation.html; Tim Hwang, *Deepfakes: A Grounded Threat Assessment* (Center for Security and Emerging Technology, July 2020), https://doi.org/10.51593/20190030; Kelly M. Sayler and Laurie A. Harris, "Deep Fakes and National Security," *Congressional Research Services*, 2022, https://crsreports.congress.gov; Luisa Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal on Selected Topics in Signal Processing* 14, no. 5 (January 2020): 910–932, ISSN: 19410484, https://doi.org/10.1109/JSTSP.2020.3002101; Hany Farid, "Creating, Using, Misusing, and Detecting Deep Fakes," *Journal of Online Trust and Safety* 1, no. 4 (September 2022), ISSN: 2770-3142, https://doi.org/10.54501/JOTS.V1I4.56.

[14] Bobby Allyn, "Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn," NPR, March 16, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[15] Rachel Metz, "How a deepfake Tom Cruise on TikTok turned into a very real AI company," CNN, August 6, 2021, https://edition.cnn.com/2021/08/06/tech/tom-cruise-deepfake-tiktok-company.

[16] This is true on two levels: first, the set of institutions that have trained their own highly capable language model from scratch has expanded rapidly over the past two years. Second, public access to many of those models has widened over time. For instance, while GPT-3 was initially released behind a sharply restricted API, it has since considerably loosened its access restrictions, allowing a larger number of people to use the model. And other, only slightly less capable models have been made fully public, with no use restrictions at all. See Section 3.2.

[17] Jennifer Earl, Thomas V. Maher, and Jennifer Pan, "The digital repression of social movements, protest, and activism: A synthetic review," *Science Advances* 8 (October 2022): 8198, https://www.science.org/doi/pdf/10.1126/sciadv.abl8198.

of future use of language models for influence operations, our focus is primarily on the technology and trends. For example, we describe tactics that could be deployed in a range of settings, rather than applications of AI to influence operations in highly specific political contexts. Additional research can expand on this paper to consider how specific groups may (or may not) use different language models for the types of influence campaigns we describe.

A paper on how current and future technological developments may impact the nature of influence operations is inherently speculative. Today, we know that it is possible to train a model and output its content—without notifying social media users—on platforms. Likewise, existing research shows that language models can produce persuasive text, including articles that survey respondents rate as credible as real news articles.[18] However, many of the future-oriented possibilities we discuss in the report are possibilities rather than inevitabilities, and we do not claim any one path will necessarily come to fruition. Similarly, our goal in this report is not to explicitly endorse any one mitigation, or any specific set of mitigations. Rather, we aim to lay out a range of possibilities that researchers and policymakers can consider in greater detail.

We also recognize that our backgrounds may result in a biased perspective: several authors work for AI developers directly, and we do not represent many of the communities that AI-enabled influence operations may affect. We encourage future research to pay particular attention to likely differential impacts and to conduct surveys of those most at risk or susceptible to AI-enabled campaigns.

1.4 Outline of the Report

The remainder of this report proceeds as follows: In Section 2, we provide an overview of influence operations, introducing key terminology, describing what influence operations are and how they are carried out, as well as providing a framework to distinguish between impact based on content and downstream impact based on trust. We focus primarily on online influence operations, in part because they are a frequent vector for text-based campaigns. In Section 3, we overview recent development in generative models and describe current access and diffusion of capabilities. In Section 4, we tie these two concepts together by examining how recent generative models could affect the future of influence operations. We describe how language models will impact the actors, behavior, and content of existing campaigns, and we highlight expected developments in the technology and critical unknowns.

The longest section of this paper is Section 5, where we move from threats to mitigations. We classify a range of potential mitigations along four key stages in the AI-to-target pipeline: model construction, model access, content dissemination, and belief formation. We conclude in Section 6 with overarching takeaways. We suggest that newer generative models have a high probability of being adopted in future influence operations, and that no reasonable mitigations can be expected to fully prevent this. However, we also suggest that a combination of multiple mitigation strategies may make an important difference and that many of these mitigations may require the formation of new collaborations between social media platforms, AI companies, government agencies, and civil society actors. In addition, we highlight several avenues for future research.

[18] Kreps, McCain, and Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation."

2 Orienting to Influence Operations

Following Russia's interference in the 2016 US election, the study of online influence operations and disinformation has grown dramatically. In this section, we begin with an overview of influence operations—what they are, why they are carried out, and the types of impacts they may (or may not) have.

2.1 What Are Influence Operations, and Why Are They Carried Out?

While there is some debate about what activities constitute an influence operation,[19] in this report, we define influence operations as *covert* or *deceptive* efforts to influence the opinions of a target audience.[20] Of note, our definition is agnostic to the truth of the message (whether the content spread is true or false) and the identity of the actor spreading it.

Influence operations include operations that intend to activate people who hold particular beliefs, to persuade an audience of a particular viewpoint, and/or to distract target audiences. The logic of distraction rests on the idea that propagandists are in competition for user attention on social media platforms, which is already spread thin.[21] If propagandists can distract target audiences from an unfavorable narrative taking shape on social media—by spreading alternative theories or diluting the information environment—they could successfully absorb user attention without necessarily persuading them.

Influence operations can come in many forms and use an array of tactics, but a few unifying themes tie many of them together. A recent report studying political influence operations in the Middle East[22] found that operations often exhibited one of several tactics:

- Attempts to cast one's own government, culture, or policies in a positive light
- Advocacy for or against specific policies
- Attempts to make allies look good and rivals look bad to third-party countries
- Attempts to destabilize foreign relations or domestic affairs in rival countries

In several of these cases, the accounts executing the operation masqueraded as locals expressing discontent with their government or certain political figures. Social media manipulation operations often employ this tactic of *digital agents of influence*, hiding the identity of the true information source from

[19] Alicia Wanless and James Pamment, "How Do You Define a Problem Like Influence?," *Journal of Information Warfare* 18, no. 3 (2019): 1–14, https://www.jstor.org/stable/26894679.

[20] Josh A. Goldstein, "Foreign Influence Operations in the Cyber Age" (PhD diss., University of Oxford, 2021), https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.840171; Ben Nimmo, *The Breakout Scale: Measuring the impact of influence operations* (Brookings Institution, September 2020), https://www.brookings.edu/research/the-breakout-scale-measuring-the-impact-of-influence-operations/.

[21] On attention economies and bounded rationality, see Elizabeth Seger et al., *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world* (The Alan Turing Institute, October 14, 2020), https://doi.org/10.17863/CAM.64183.

[22] M.A. et al., "Middle East Influence Operations: Observations Across Social Media Takedowns," *Project on Middle East Political Science*, August 2021, https://pomeps.org/middle-east-influence-operations-observations-across-social-media-takedowns.

the target audience.[23] Russia's Internet Research Agency (IRA) accounts, for example, pretended to be Black Americans and conservative American activists, and directly messaged members of each targeted community. Identifying these inauthentic accounts often relies on subtle cues: a misused idiom, a repeated grammatical error, or even the use of a backtick (`) where an authentic speaker would use an apostrophe (‘). State-level adversarial actors often run a combination of tactics, leveraging their own employees or outsourcing to digital mercenaries.

Since 2016, Meta and Twitter have removed well over a hundred social media influence operations, stemming from dozens of different countries.[24] These operations often include persona creation (creating fake identities to spread a message), fake news properties, and inauthentic amplification efforts. But influence operations have also expanded significantly beyond Facebook and Twitter and into alternative platforms, small group settings, and encrypted spaces.[25] Reporting from the *New York Times*, in hand with Israeli disinformation researchers, documented how "Iranian agents had infiltrated small [Israeli] WhatsApp groups, Telegram channels and messaging apps" to spread polarizing content.[26] At times these influence operations display novel ingenuity, leveraging platform policies in an adversarial fashion. A campaign supporting the Tanzanian government that was removed by Twitter in 2021, for example, used false claims of copyright reporting to target Tanzanian activists' accounts.[27]

Much of the recent research and public attention on influence operations focuses on foreign campaigns—where governments or citizens in one country target citizens in a different country.[28] But, as the Tanzania example shows, influence operations can also be domestically focused. Political actors frequently spread covert propaganda targeting their citizens in order to boost their popularity, undermine that of an opponent, or sow confusion in the political system. In 2020, Facebook suspended fake personas spreading polarizing content about Brazilian politics that were linked to Brazilian lawmakers as well as President Jair Bolsonaro and his sons, Congressman Eduardo Bolsonaro and Senator Flavio Bolsonaro.[29] In fact,

[23] Russia, for example, leverages personas that speak as if they are members of the targeted communities. Some of the personas produce short-form content, such as tweets and Facebook posts. Others masquerade as journalists and write long-form narrative content that they then submit to legitimate publications or publish on Russian self-administered proxy "media outlets" or "think tanks." For examples in the Russia context, see Renee DiResta and Shelby Grossman, *Potemkin Pages & Personas: Assessing GRU Online Operations, 2014-2019* (Stanford Internet Observatory, 2019), https://cyber.fsi.stanford.edu/io/publication/potemkin-think-tanks. For another example, see 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. For a variant of this tactic leveraging compromised websites, see Mandiant, *'Ghostwriter' Influence Campaign: Unknown Actors Leverage Website Compromises and Fabricated Content to Push Narratives Aligned with Russian Security Interests* (Mandiant), https://www.fireeye.com/content/dam/fireeye-www/blog/pdfs/Ghostwriter-Influence-Campaign.pdf. For examples of front proxy media sites and "think tanks," see *Pillars of Russia's Disinformation and Propaganda Ecosystem* (U.S. Department of State, August 2020), https://www.state.gov/russias-pillars-of-disinformation-and-propaganda-report/

[24] Disinfodex (August 2020), database distributed by Carnegie Endowment for International Peace, https://disinfodex.org/; Gleicher et al., *Threat Report: The State of Influence Operations 2017-2020*. Note, these are only the operations that have been found and publicly reported. Because influence operations are typically designed to be kept secret, it likely reflects an undercount of all operations on these platforms.

[25] Graphika, *Posing as Patriots* (Graphika, June 2021), https://graphika.com/reports/posing-as-patriots.

[26] Sheera Frenkel, "Iranian Disinformation Effort Went Small to Stay Under Big Tech's Radar," *New York Times*, June 30, 2021, https://www.nytimes.com/2021/06/30/technology/disinformation-message-apps.html.

[27] Shelby Grossman et al., "The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists," Stanford Internet Observatory, December 2, 2021, https://stacks.stanford.edu/file/druid:bt877dz8024/20211202-tz-twitter-takedown.pdf.

[28] Claire Wardle, "The Media Has Overcorrected on Foreign Influence," *Lawfare*, October 26, 2020, https://www.lawfareblog.com/media-has-overcorrected-foreign-influence.

[29] Jack Stubbs and Joseph Menn, "Facebook suspends disinformation network tied to staff of Brazil's Bolsonaro," *Reuters*, July 8, 2020, https://www.reuters.com/article/us-facebook-disinformation-brazil/facebook-suspends-disinformation-network-tied-to-staff-of-brazils-bolsonaro-idUSKBN2492Y5.

many commentators believe that *domestic*, not foreign, influence operations are the most worrisome.[30] Influence operations have additionally been deployed to take sides in intraparty politics,[31] and, in the case of several attributed to the Chinese Communist Party, to target diaspora populations.[32]

2.2 Influence Operations and Impact

Influence operations can have impact based on their specific content or focus (e.g., through persuasion), or by eroding community trust in the information environment overall.

In current influence operations, direct impact from content is sometimes limited by resources, quality of the message, and detectability of the operation. These factors may matter differently depending on the goals of the operator—for instance, if operators are looking only to distract instead of to convince targets of a specific viewpoint, the quality of each individual message is likely far less significant. In theory, however, these constraints may be partially overcome by language models in the future.

Having an effect on trust in an information environment depends less on the substance and more on creating the perception that any given message might be inauthentic or manipulative. Even if influence operations do not change someone's views, they may lead people to question whether the content they see from even credible sources is in fact real, potentially undermining faith in democratic and epistemic institutions more broadly.

2.2.1 Impact Based on Content

An influence operation could have impact based on content if it (1) persuades someone of a particular viewpoint or reinforces an existing one, (2) distracts them from finding or developing other ideas, or (3) distracts them from carving out space for higher quality thought at all. Often the goal is simply to distract from information that is potentially harmful to the operator.[33] As advertisers, media outlets, and platforms already compete for viewers, distraction operations can often exploit and exacerbate such preexisting attention competitions to crowd out important information with attention-grabbing, irrelevant information. Distraction operations therefore do not require a target to be persuaded by the information spread, but rather that a target not be persuaded by (or even consider) some other piece of information.

There are both historical and contemporary examples where the impact of an influence operation can be clearly measured or traced. For example, in the 1980s during the HIV epidemic, the Soviet Union waged an influence operation spreading the claim that the United States government created the virus

[30] Emerson T. Brooking and Jacob Shapiro, "Americans Were Worried About the Wrong Threat," Atlantic, January 10, 2020, https://www.theatlantic.com/ideas/archive/2021/01/bigger-threat-was-always-domestic/617618/.

[31] Shelby Grossman et al., *Staying Current: An Investigation Into a Suspended Facebook Network Supporting the Leader of the Palestinian Democratic Reform Current* (Stanford Internet Observatory, February 10, 2021), https://purl.stanford.edu/tk756wp5109.

[32] "Chinese propagandists court South-East Asia's Chinese diaspora," Economist, November 20, 2021, https://www.economist.com/asia/2021/11/20/chinese-propagandists-court-south-east-asias-chinese-diaspora.

[33] Gary King, Jennifer Pan, and Margaret E. Roberts, "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument," *American Political Science Review* 111, no. 3 (2017): 484–501, https://doi.org/10.1017/S0003055417000144.

in a lab. One 2005 study found that 27% of African Americans still believed this claim.[34] In 2016, the IRA used manipulative agents of influence on Facebook to provoke real-world conflict by organizing protests and counter-protests outside the Islamic Da'wah Center in Houston.[35] The impact is relatively easy to trace here because the protests would not have occurred without the IRA's activity. A recent literature review examining social science research on the effects of influence operations found "strong evidence that long-term campaigns on mass media have measurable effects on beliefs and consequential behaviors such as voting and risk-taking combat." While noting that evidence remains sparse, the study also found there is "some evidence that social media activity by exceptionally influential individuals and organizations can stoke low-level violence."[36]

However, the impact and effectiveness of influence operations are usually difficult to measure. Disinformation researchers typically focus on engagement metrics—things like clicks and shares—which are inadequate proxy measures of social influence.[37] In cases where a clear comparison group does not exist, it can be difficult to determine how viewing or engaging with content translates into important political outcomes like polarization or votes. While platforms make attributions and provide researchers with data about taken-down influence operations, researchers still have limited visibility into the impact on users or their subsequent behavior after engagement. Furthermore, not all influence operations are detected. Even propagandists who attempt to measure their own impact can face challenges given multicausality and difficulties in measuring opinion change over time. As scholars have noted, this ambiguity has historically contributed to intelligence agencies inflating the impact of their influence operations for bureaucratic gain.[38]

Despite these measurement challenges, some features clearly limit the impact of existing campaigns, including resources, content quality and messaging, and detectability. We outline these limitations below, and discuss in the following section how generative models may help overcome these barriers.

- **Resources:** Like marketing campaigns, the success of an influence operation is a function of resources and the ability to get the desired content in front of one's target. How many propagandists does a political actor hire to write content? How many social media accounts can they obtain to fake popularity? Low-resourced campaigns are less likely to get their desired content in front of

[34] Renee DiResta, Michael McFaul, and Alex Stamos, "Here's How Russia Will Attack the 2020 Election. We're Still Not Ready.," *The Washington Post*, November 15, 2019, https://www.washingtonpost.com/opinions/2019/11/15/heres-how-russia-will-attack-election-were-still-not-ready/.

[35] Martin J. Riedl et al., "Reverse-engineering political protest: the Russian Internet Research Agency in the Heart of Texas," *Information, Communication, and Society* 25, no. 15 (2021), ISSN: 14684462, https://doi.org/10.1080/1369118X.2021.1934066.

[36] John 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.

[37] For example, researchers conducted a study comparing Twitter users who interacted with content from the IRA with those who did not. The study found "no substantial effects of interacting with Russian IRA accounts on the affective attitudes of Democrats and Republicans who use Twitter frequently toward each other, their opinions about substantial political issues, or their engagement with politics on Twitter in late 2017." Christopher A. Bail et al., "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017," *PNAS* 117, no. 1 (January 7, 2020), https://doi.org/10.1073/pnas.1906420116

[38] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York: Farrar, Straus, Giroux, 2020), 260, https://us.macmillan.com/books/9780374287269/activemeasures.

the target or to garner media coverage.[39]

- **Quality and Message of Content:** People are less likely to be persuaded by messaging if it strongly counters their established attitude or if the arguments are poorly constructed or poorly reasoned.[40] Campaigns with messaging that disconfirms targets' attitudes, does not successfully blend in with a target's information environment, and provides low-quality arguments are, all else being equal, less likely to be effective.[41]

- **Detectability:** Finally, operations that are quickly discovered are less likely to have an impact. Social media platforms and independent researchers actively search for influence operations, and platforms remove them in order to limit their reach. In fact, awareness that these operations may be removed can itself shape the behavior of propagandists, leading them to pursue distraction operations if they believe persona development—which requires longer-term investment but can be more persuasive to observers—is not worth the effort.[42]

It is helpful to keep these limitations in mind as we consider the role that language models can play in influence campaigns. If they can overcome existing limitations, then they may pose a significant issue for the information environment. We discuss this further in Section 4.

2.2.2 Downstream Impact Based on Trust

The second way that influence operations can have an impact is by eroding trust. Degrading societal trust does not necessarily require high quality efforts: even when influence campaigns are detected, their appearance, especially at scale, may cause users to become suspicious of other, authentic sources.[43] Propagandists often aim to exploit vulnerabilities in their target's mental shortcuts for establishing trust, especially where information technologies make it harder to evaluate the trustworthiness of sources. By manipulating public perceptions of reputation, harnessing fake or misleading credentials and testimonials, or tampering with photographic and video evidence, influence operators can serve to undermine

[39] Beyond simply expanding the size of a campaign, greater resources may help operators target their content to a wider range of people. Research on the 2016 election suggests that fake news consumption was heavily concentrated, with only 1% of Twitter users exposed to 80% of fake news. Nir Grinberg et al., "Fake news on Twitter during the 2016 U.S. presidential election," *Science* 363, no. 6425 (January 25, 2019): 374–378, ISSN: 10959203, https://doi.org/10.1126/science.aau2706

[40] Hee Sun Park et al., "The Effects of Argument Quality and Involvement Type on Attitude Formation and Attitude Change: A Test of Dual-Process and Social Judgment Predictions," *Human Communication Research* 33, no. 1 (January 2007): 81–102, ISSN: 0360-3989, https://doi.org/10.1111/J.1468-2958.2007.00290.X.

[41] However, as discussed above, note that the importance of this factor depends on the goals of the operator. If the goal is pure distraction, having high-quality posts may be far less significant than if the operator is aiming to actually persuade.

[42] Josh A. Goldstein and Renee DiResta, "China's Fake Twitter Accounts Are Tweeting Into the Void," *Foreign Policy*, December 15, 2021, https://foreignpolicy.com/2021/12/15/china-twitter-trolls-ccp-influence-operations-astroturfing/. We recognize that, in some cases, influence operators desire their efforts to be detected in order to stir worry among a target population. However, because many influence operations seek to directly change opinions, and universally easy detection would undermine efforts to stir worry, we treat lower detectability as desirable to propagandists.

[43] Recent research suggests that educating people about deepfakes makes them more likely to believe that real videos they subsequently see are also fakes; see John Ternovski, Joshua Kalla, and Peter Aronow, "The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments," *Journal of Online Trust and Safety* 1, no. 2 (February 2022), ISSN: 2770-3142, https://doi.org/10.54501/JOTS.V1I2.28. Politicians may also benefit from the "liar's dividend" by falsely claiming that real events that paint them in a critical light are fake news or deepfakes. See Robert Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (2019): 1753, https://doi.org/10.15779/Z38RV0D15J.

trust beyond the specific topic of their campaign.[44] Lower societal trust can reduce a society's ability to coordinate timely responses to crises, which may be a worthy goal for adversarial actors in and of itself.

In turn, lower societal trust also creates a more favorable operating environment for propagandists to pursue their objectives. Preexisting polarization and fragmentation in society undercut the ability of honest actors to establish broad credibility, and can give influence operators a foothold to tailor their messaging to narrower audiences, sow division, and degrade social capital and institutional trust. Low general trust undermines the norms that enable people and organizations to interact and cooperate without extensive rules and processes to govern their behavior.[45]

[44] Seger et al., *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*.

[45] Lower societal trust also increases transaction costs. In the economy, this decreases the efficiency of markets, and in government, it incentivizes regulatory overreach and accordingly bureaucratic growth that can entrench interests and degrade institutional agility. See Michael J. Mazarr et al., *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (RAND Corporation, October 2019), 62, https://doi.org/10.7249/RR2714.

3 Recent Progress in Generative Models

Understanding the present state of generative models is helpful for addressing their potential role in influence operations. This section introduces generative models to disinformation researchers and policymakers, and will likely be familiar to those in the machine learning (ML) community.

3.1 What Are Generative Models, and How Are They Built?

In the last decade, research in AI has improved the ability to automate the production of digital content, including images, video, audio, and text. These new generative AI models can learn to understand the patterns in a given type of data—like text in the English language or the audio waveforms comprising songs—in order to sample new items of that type and produce original outputs. In a wide number of domains, progress in generative models over the past decade has moved shockingly quickly and produced surprisingly realistic output, as illustrated in Table 3 and Figures 1, 2, and 3.

2011	2020
The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger	**The meaning of life** is contained in every single expression of life. It is present in the infinity of forms and phenomena that exist in all aspects of the universe.

Table 3: Generative text model outputs in 2011 versus 2020.[46]

These machine language systems consist of large artificial neural networks[47] and are "trained" via a trial-and-error process over mountains of data.[48] The neural networks are rewarded when their algorithmically generated words or images resemble the next word in a text document or a face from an image dataset.[49] The hope is that after many rounds of trial and error, the systems will have picked up general features of the data they are trained on. After training, these generative models can be repurposed to generate entirely new synthetic artifacts.

[46] The 2011 text was generated from Ilya Sutskever, James Martens, and Geoffrey Hinton, "Generating Text with Recurrent Neural Networks," ed. Lisa Gooter and Tobias Scheffer, *Proceedings of the 28th International Conference on Machine Learning*, 2011, https://icml.cc/2011/papers/524_icmlpaper.pdf. The 2020 text was generated using the 175B GPT-3 model.

[47] Artificial neural networks are a class of statistical models that are loosely inspired by biological brains. For a technical introduction discussing the role of neural networks in modern machine learning, see the Introduction in Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (MIT Press, 2016), https://www.deeplearningbook.org/. For an introduction for policymakers, see Ben Buchanan and Taylor Miller, *Machine Learning for Policy Makers: What It Is and Why It Matters* (Belfer Center for Science and International Affairs, June 2017), https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

[48] For example, HuggingFace's BigScience project is using a training dataset of 1.5 TB (see "Building a TB Scale Multilingual Dataset for Language Modeling," Hugging Face BigScience, https://bigscience.huggingface.co/blog/building-a-tb-scale-multilingual-dataset-for-language-modeling); the original GPT-3 project (published in 2021) used a filtered dataset of 570 GB; the largest DeepMind's Gopher model saw about 1.3 TB of text. The text is composed via sources like web crawls, Wikipedia, scanned books, and news articles.

[49] Other methods to train generative models are also in development. For example, diffusion models have been applied to text-to-image generation; see Aditya Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arxiv:2204.06125 [cs.CV]*, April 2022, https://doi.org/10.48550/arxiv.2204.06125.

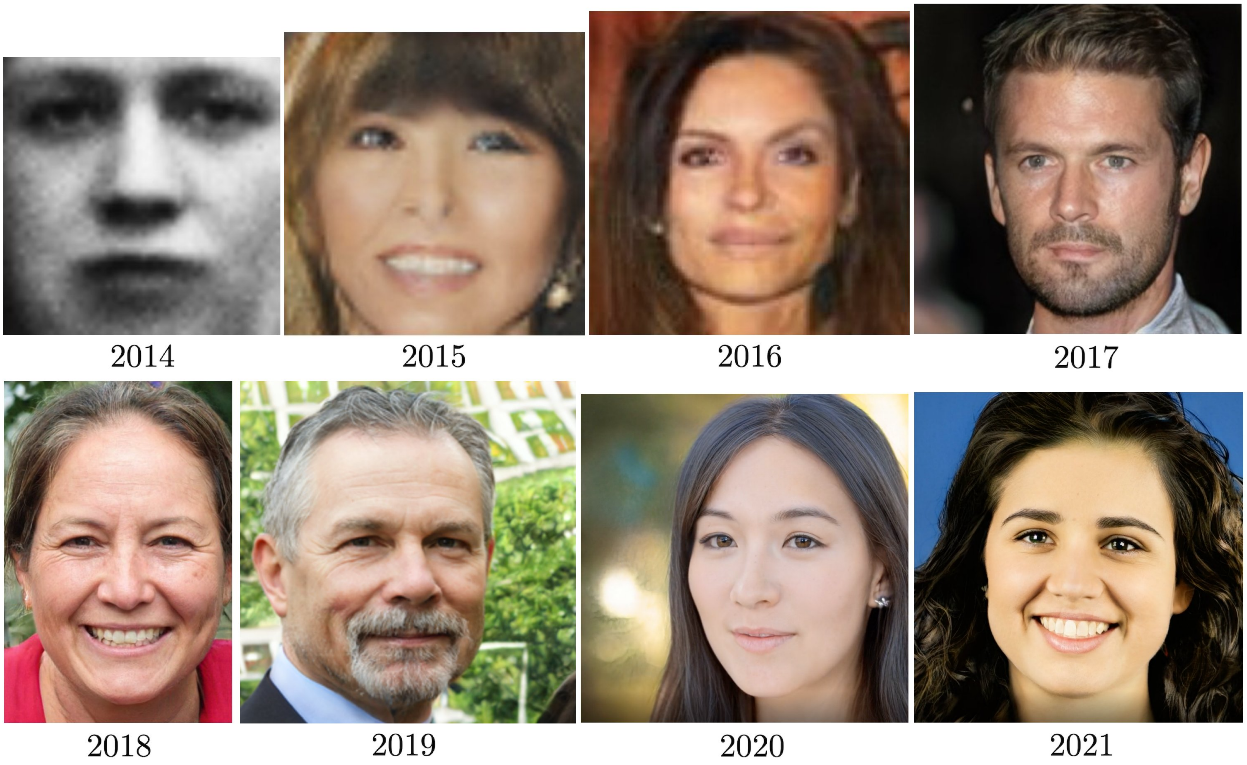
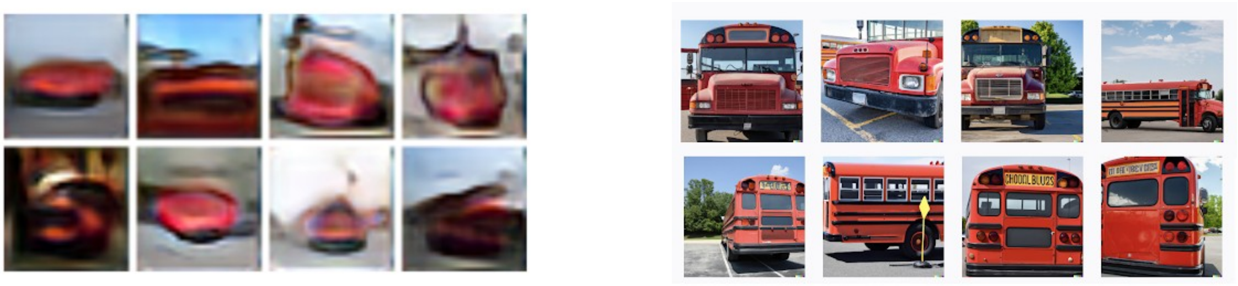


Figure 1: Seven years of progress in synthetic face generation. All of these images are produced with Generative Adversarial Networks.[51]



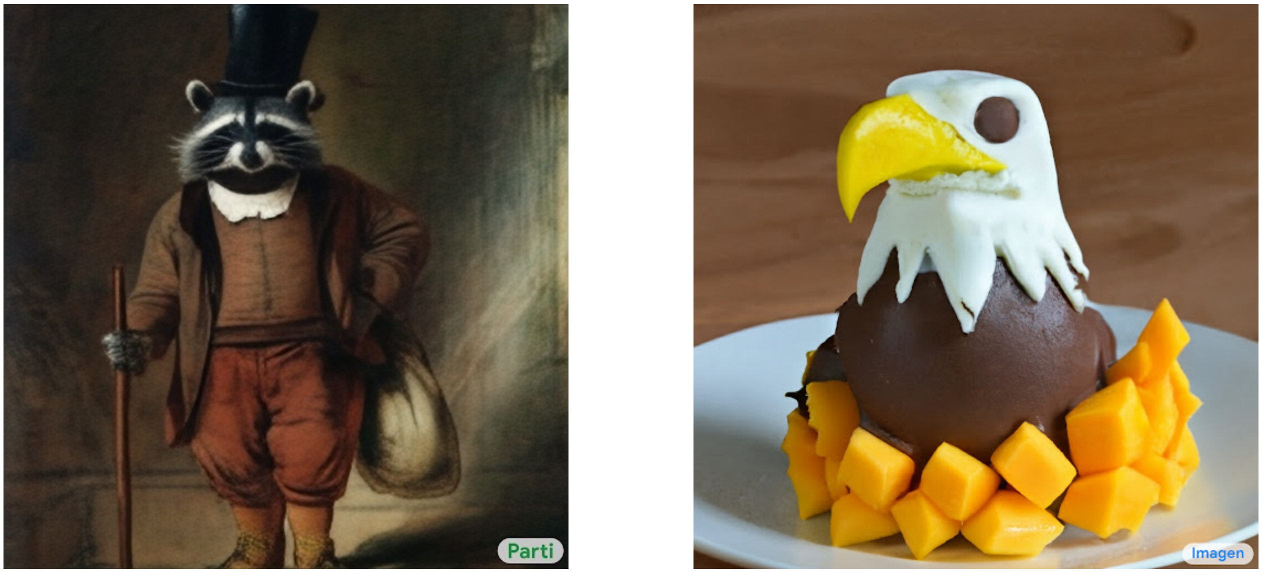
(a) 2015

(b) 2022

Figure 2: Seven years of progress in image generation from language. Left image from a 2015 paper, which introduced one of the first methods to generate images from text. The prompt is taken from that paper and intends to show novel scenes. On the right, the same prompt is run on OpenAI's DALL•E 2. Today's systems can easily do certain tasks that were hard in 2015.[52]

[51] Original source: Tamay Besiroglu (@tamaybes), "7.5 years of GAN progress on face generation," Twitter, October 20, 2021, 10:15 AM, https://twitter.com/tamaybes/status/1450873331054383104, building on Ian Goodfellow, (@goodfellow_ian), Twitter, January 14, 2019, 4:40 PM, https://twitter.com/goodfellow_ian/status/1084973596236144640.

[52] Elman Mansimov et al., "Generating Images from Captions with Attention," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, November 9, 2015, https://doi.org/10.48550/arxiv.1511.02793.



(a) *"A raccoon wearing formal clothes, wearing a top hat and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of Rembrandt"*

(b) *"A bald eagle made of chocolate powder, mango, and whipped cream"*

Figure 3: Elaborate scene construction and composition with 2022 text-to-image models. While Figure 2 shows that 2022 models can do hard tasks from 2015 easily, text-to-image models can also do tasks that were not possible before. In this image, many details of the scene are described via language, and the system translates that into a plausible image. Left is from Google's Parti, and right is from Google's Imagen.[53]

Creating generative models from scratch involves two steps. The first is to take a neural network and train it on an immense amount of raw data. This training process automatically adjusts the many (sometimes more than hundreds of billions) "parameters" of the neural network, which are somewhat analogous to synapses in biological brains. This step culminates in a system that is quite general (it can do many different tasks) and capable (it can do these tasks well),[50] but that may be difficult to use for specific tasks or that may still lack certain specialized skills. The optional second—and much cheaper—step is to refine this foundation model by further training (or "fine-tuning") it on small amounts of task-specific data. Fine-tuning can extend a model's capabilities—for example, a model can be fine-tuned to imitate complex human behaviors like following instructions—or it can be used to train domain-specific skills in smaller models.

Training a state-of-the-art, large generative model from scratch in 2022 can involve costs that are at least tens of millions of dollars.[54] However, it is becoming less expensive to reach near state-of-the-art performance: while it originally cost millions of dollars to train GPT-3 in 2020, in 2022 MosaicML was able to train a model from scratch to reach GPT-3 level performance for less than $500k.[55] Because of

[53] Jiahui Yu et al., "Parti: Pathways Autoregressive Text-to-Image Model," https://parti.research.google/; Chitwan Saharia et al., "Imagen: Text-to-Image Diffusion Models," https://imagen.research.google/.

[50] Bommasani et al., "On the Opportunities and Risks of Foundation Models."

[54] An estimate for Google's PaLM model puts it at ~$23M; see Lennart Heim, "Estimating PaLM's training cost," .xyz Blog, April 5, 2022, https://blog.heim.xyz/palm-training-cost/. Estimates for other language models are also in the single-to-double-digit millions of dollars.

[55] Abhinav Venigalla and Linden Li, "Mosaic LLMs (Part 2): GPT-3 quality for <$500k," Mosaic, September 29, 2022, https://www.mosaicml.com/blog/gpt-3-quality-for-500k.

this upfront cost, many developers will choose to fine-tune an existing model for their task. This allows them to leverage the general capabilities of the foundation model—imbued from pre-training—at lower cost.[56]

Recent advances in generative models have been driven by three major developments: (1) the explosion of training data in the form of human language available on the internet (and in curated datasets of internet or user-generated content); (2) improvements in the underlying neural network models and the algorithms used to train them; and (3) rapid growth in the amount of computational power that leading actors have used to train these models, which allows for the creation of larger, more sophisticated models. In many cutting-edge applications, acquiring sufficient computational power to train a model is the most expensive of these components, and the relative capability of different models tends to roughly correspond to how much computational power was used to train them.[57]

Requirements to Create a Cutting-Edge Language Model	**Cause of Recent Improvement**
Data	Explosion of available training data (text on the internet)
Algorithm	Improvements in large-scale training algorithms and neural network architectures
Computational Power (compute)	Increase in availability of computational power for AI scientists and improvements in methods to leverage that compute

Table 4: Summary of Training Requirements and Areas of Recent Improvement of Language Models

Generative language models that "understand" and produce language are the central focus of this report.[58] In principle, a system that can receive and output arbitrary text can perform every task that is expressible via text. Interacting with a language model is, in some sense, like interacting with a remote employee over a textual interface. While current language models are not nearly at human level, they have made great strides in their generality and capability[59]. For example, not only can the same system (the hypothetical "employee") carry out the task of classifying tweets as positive or negative sentiment, but it can also generate tweets, write summaries, carry on conversations, write rudimentary source code, and so on.[60]

While impressive, current generative language models have many limitations. Even the most sophisticated systems struggle to maintain coherence over long passages, have a tendency to make up false or absurd statements of fact, and are limited to a generation length of about 1,500 words. In addition,

[56] Sebastian Ruder, "Recent Advances in Language Model Fine-tuning," Sebastian Ruder (Blog), February 24, 2021, https://ruder.io/recent-advances-lm-fine-tuning/.

[57] For elaboration on these points, see Deep Ganguli et al., "Predictability and Surprise in Large Generative Models," *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 2022, 1747–1764, https://doi.org/10.1145/3531146.3533229.

[58] Other generative models may focus on generating and modeling visual information—as in images or video—or audio information. In principle, generative models may model any type of sensory information. For a review of audio models, see Zhaoxi Mu, Xinyu Yang, and Yizhuo Dong, "Review of end-to-end speech synthesis technology based on deep learning," *arxiv:2104.09995 [cs.SD]*, April 2021, https://doi.org/10.48550/arxiv.2104.09995. For an example of a video model, see Emmanuel Kahembwe and Subramanian Ramamoorthy, "Lower Dimensional Kernels for Video Discriminators," *Neural Networks* 132 (December 2020): 506–520, https://doi.org/10.1016/j.neunet.2020.09.016.

[59] We describe future developments of these dimensions of progress in Section 4.2.2.

[60] See Google's PaLM system for some examples: Sharan Narang and Aakanksha Chowdhery, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," Google AI Blog, April 5, 2022, https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html.

models perform worse as they are given more cognitively complex tasks: for instance, asking a generative model to write a few conservative-leaning tweets on a topic will likely result in better outputs than asking a model to rewrite an existing news story in a way that subtly promotes a conservative narrative.[61] While these limitations are noteworthy, progress in generative models is both rapid and hard to predict. The capabilities of current models should be considered lower bounds on how realistic generative model outputs can become, and it is not clear where the relevant upper bound is—if it exists.

To overcome these limitations, ongoing research targets improvements in data, algorithms, and computational power. For example, some research attempts to improve the quality of the data that the neural network ingests. One way to do so is by collecting data from human domain experts or demonstrators of the desired capability.[62] Improvements in neural network architectures and new training strategies to imbue the model with improved capability can lead to better algorithms. And, of course, training models on more powerful supercomputers increases the amount of computational power available to the model.

3.2 Access and Diffusion of Generative Models

A sizable number of organizations have developed advanced language models. These models are accessible on a spectrum from fully public to fully private. A small number of models are fully public, meaning that anyone can download and use them to produce outputs in a way that can no longer be monitored by the models' designers. The largest openly downloadable model as of September 2022 (measured by the number of parameters in the neural network model) is BLOOM by HuggingFace's BigScience project—a 175 billion- parameter model openly released in July 2022. However, algorithmic improvements have also enabled much smaller open source models that rival or exceed BLOOM and GPT-3 on several capabilities.[63]

Other models have been kept fully private, with no means for non-developers to access or use the model. DeepMind's Gopher (280 billion parameters) and Microsoft and Nvidia's Megatron-Turing NLG (530 billion parameters, but not fully trained)—both of which were created primarily for research purposes—fall into this category. As mentioned previously, the relative capabilities of different language models tends to correspond to the amount of computational power used to train them, and more computational power generally (though not always) means a larger model with more parameters.[64] It is therefore worth emphasizing that the largest fully public model is two to three times smaller than the largest currently existing private models. However, this may change soon if more developers open-source their models or a model is leaked.

[61] Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation*.

[62] For example, to train models to play Minecraft, researchers collected demonstrations of behaviors from humans; see Bowen Baker et al., "Learning to Play Minecraft with Video PreTraining (VPT)," OpenAI Blog, June 23, 2022, https://openai.com/blog/vpt/. A survey with more examples is available in Xingjiao Wu et al., "A Survey of Human-in-the-loop for Machine Learning," *Future Generation Computer Systems* 135 (August 2021): 364–381, https://doi.org/10.1016/j.future.2022.05.014.

[63] Hyung Won Chung et al., "Scaling Instruction-Finetuned Language Models," *arxiv:2210.11416 [cs.LG]*, October 20, 2022, https://doi.org/10.48550/arxiv.2210.11416.

[64] Advances in sparsity and retrieval methods are two ways that the number of parameters can come apart from both the computational power used to train the model and the model's capabilities. See Noam Shazeer et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, January 2017, https://doi.org/10.48550/arxiv.1701.06538; Sebastian Borgeaud et al., "Improving language models by retrieving from trillions of tokens," *arxiv:2112.04426 [cs.CL]*, December 2021, https://doi.org/10.48550/arxiv.2112.04426.

A third category of models attempt to balance public and private access. Meta AI gave some external researchers copies of its 175 billion-parameter language model while requiring them to sign a license that banned certain use cases.[65] Another method allows for users to sign up for certain types of access through an application programming interface (API). An API-based access regime allows AI developers to commercialize access to their model, track model usage, and impose restrictions on both who can access the model and how they can use it. GPT-3, Jurassic-1, and Cohere Extremely Large, for instance, are all currently accessible via an API.[66] Keeping models behind an API allows developers a great deal of discretion regarding the conditions under which their model can be accessed.[67] Organizations that use an API-based access regime ensure that users can submit queries to a model and receive outputs, but also that users cannot directly see or download the model itself,[68] which means that they cannot fine-tune it for their own specific applications. An AI provider may also choose to support API-based fine-tuning, which would allow the AI developer to monitor and restrict certain fine-tuning use cases.[69]

Table 5 includes an illustrative list of the most capable current (publicly known, as of September 2022) language models that vary across access regime, primary language of output, and sizes. There are several key takeaways that characterize the current state of model diffusion.

First, anyone can access a number of moderately capable models that have been made fully public, but the most capable models remain either private or kept behind monitorable APIs. While currently publicly available models may not be as powerful as the largest private models, they can likely be fine-tuned to perform remarkably well on specific tasks at far less cost than training a large model from scratch. This type of fine-tuning might not be within the reach of most individuals, but it is likely feasible for any nation-state as well as many non-state actors, such as firms and wealthy individuals.[71]

Second, in addition to cutting-edge models from AI developers like Google (US) and DeepMind (UK), several international actors have developed highly capable models likely motivated by commercial interests and as a matter of national prestige. For example, Inspur's Yuan 1.0, a 245 billion-parameter Chinese-language model, and Naver's HyperClova, a 204 billion-parameter Korean-language model,

[65] Including "military purposes" and "purposes of surveillance"; see "OPT-175B License Agreement," Metaseq, https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/MODEL_LICENSE.md.

[66] "API," OpenAI, accessed January 31, 2022, https://openai.com/api/; Kyle Wiggers, "Announcing AI21 Studio and Jurassic-1 Language Models," AI21 Labs, accessed January 31, 2022, https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1; Cohere, "About," accessed January 31, 2022, https://docs.cohere.ai/api-reference/.

[67] However, because external researchers do not have access to the raw models from these APIs, API-based access regimes may make it more difficult for researchers to replicate and improve the private models.

[68] API-based models may not be immune to manipulation or theft by adversaries. Model inversion attacks can allow an adversary to potentially steal a model by querying an API many times; see Florian Tramer et al., "Stealing Machine Learning Models via Prediction APIs," *25th USENIX Security Symposium (Austin, TX; USENIX Security 16)*, 2016, 601–618, https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer. However, these methods are expensive and have not been demonstrated to work in practice against a foundation model API.

[69] For example, Cohere and OpenAI offer fine-tuning through their APIs: "Finetuning Generation Models," Cohere, accessed June 2022, http://web.archive.org/web/20220621204451/https://docs.cohere.ai/finetuning-wiki/; "Fine-tuning," OpenAI, accessed June 2022, https://beta.openai.com/docs/guides/fine-tuning

[70] Model sizes come from Jaime Sevilla et al., "Compute Trends Across Three Eras of Machine Learning," *Proceedings of the International Joint Conference on Neural Networks*, March 9, 2022, https://doi.org/10.48550/arxiv.2202.05924; Jaime Sevilla et al., "Parameter, Compute and Data Trends in Machine Learning," 2021, https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=0; and Jeffrey Ding and Jenny Xiao, "Recent Trends in China's Large-Scale Pre-Trained AI Models," *(Working Paper)*. Yalm-100B's compute usage is estimated assuming use of a GPT model in full precision for 300B tokens; see Mikhail Khrushchev, "Yandex Publishes YaLM 100B. It's the Largest GPT-Like Neural Network in Open Source," Medium, June 23, 2022, https://medium.com/yandex/yandex-publishes-yalm-100b-its-the-largest-gpt-like-neural-network-in-open-source-d1df53d0e9a6.

[71] Furthermore, as mentioned above, some AIaaS providers offer fine-tuning as a service.

Model	Size: Training Computation (PFLOP)[70]	Size: Parameters	Organization	Date of Announcement	Primary Language	Access Regime	Resource
Ernie 3.0 Titan	4.2×10^7	260B	Baidu	Dec 2021	Chinese	Restricted (API)	Outputs
Pan-Gu-alpha	5.80×10^7	200B	Huawei	Apr 2021	Chinese	Private	-
Hyper-CLOVA	6.30×10^7	204B	Naver Corp.	Sep 2021	Korean	Private	-
GPT-NeoX	9.30×10^7	20B	Eleuther AI	Feb 2022	English	Public	Parameters
Yalm-100B	1.80×10^8	100B	Yandex	Jun 2022	Russian	Public	Parameters
GPT-3	3.00×10^8	175B	OpenAI	May 2020	English	Restricted (API)	Outputs
Yuan 1.0	4.10×10^8	245B	Inspur	Oct 2021	Chinese	Restricted (API)	Outputs
OPT-175B	4.30×10^8	175B	Meta	Jan 2022	English	Restricted (license)	Parameters
BLOOM	6.04×10^8	175B	BigScience	July 2022	Multiple	Public	Parameters
Gopher	6.30×10^8	280B	DeepMind	Dec 2021	English	Private	-
Megatron-Turing	1.40×10^9	530B	Microsoft, NVIDIA	Jan 2022	English	Private	-
PaLM	2.50×10^9	540B	Google	Apr 2022	English	Private	-

Note: We order the table by training computation requirements as a proxy for capability.

Table 5: Illustrative List of State-of-the-Art Language Models.

have matched and exceeded the size of GPT-3 and likely offer similarly impressive capabilities.[72] While access to PanGu-α, HyperClova, and Wu Dao 2.0 looks likely to remain partially or fully restricted, other models are public. For example, the Russian Yalm 100 billion-parameter model is openly available through code repositories on GitHub and/or HuggingFace.[73] Some of the Beijing Academy of Artificial Intelligence's (BAAI) WuDao models are directly downloadable from their website.[74]

Third, these international actors have optimized their models for their national languages. For example, the Yuan 1.0 model excels in Chinese-language tasks. While per-language performance can be approximated by the proportion of training data that is in a particular language, models can also perform well at producing text in multiple languages or translating between them—if the model is trained on enough data from multiple languages. This trend of language-specific optimization suggests that if these models are applied to influence operations, they will be most able to target populations speaking specific languages that are well-represented in a particular model's training data.

[72] See Wei Zeng et al., "PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation," *arxiv:2104.12369 [cs.CL]*, April 2021, https://doi.org/10.48550/arxiv.2104.12369; Kyle Wiggers, "Huawei trained the Chinese-language equivalent of GPT-3," VentureBeat, April 29, 2021, https://venturebeat.com/ai/huawei-trained-the-chinese-language-equivalent-of-gpt-3/; "NAVER Unveils HyperCLOVA, Korea's First Hyperscale 'Al to Empower Everyone'," *Naver Corp. Press Releases*, May 25, 2021, https://www.navercorp.com/en/promotion/pressReleasesView/30686.

[73] For example: "Muse API," PAGnol, https://muse.lighton.ai/home; Anton Emelyanov et al., "Russian GPT-3 models," GitHub, https://github.com/ai-forever/ru-gpts#readme.

[74] "WudaoAI," *Beijing Academy of Artificial Intelligence*, accessed October 30, 2022, https://wudaoai.cn/model/.

4 Generative Models and Influence Operations

This section marries the previous sections' emphases on influence operations and generative models. We build on the existing but nascent body of research on AI-generated influence campaigns in two steps. First, we introduce the ABC framework—actors, behaviors, and content—that is well-known among disinformation researchers, and describe how generative models may transform each of these three facets.[75] Then, we examine expected developments and critical unknowns in the field of machine learning that will impact the role that generative models can play in influence operations. For each expected development, we describe the current state of technology, expected improvements, and the implications such improvements would have for the future of influence campaigns.

4.1 Language Models and the ABCs of Disinformation

In this paper, we build on the "ABC" model, a popular model in the disinformation field, that distinguishes between key manipulation vectors in disinformation campaigns.[76] "A," for **actors**, references the fact that the entity behind a campaign is often not what it seems; for example, the accounts in a conversation may look like Black Lives Matter activists, but in reality may be state-linked actors using fake accounts in active misdirection. "B" is for **behavior**, and refers to *how* propagandists wage their campaigns—the techniques used to perpetuate disinformation, such as the use of automation or attempts to manipulate engagement statistics via click farms.[77] "C" alludes to the **content** itself, the substance (narrative, memes, etc.) that the accounts are attempting to launder or amplify; this third facet of disinformation campaigns is perhaps the most visible to the public, and media will highlight the substance in its coverage.[78] Although, as discussed in the Section 1, we are focused on influence operations, not disinformation exclusively, this model helps characterize potential changes that may arise due to language models.

One of the reasons that platforms and researchers assess all three dimensions—the actors, behaviors, and content—when evaluating an influence operation is that at times one facet may be perfectly authentic even within an overall manipulative campaign. Authentic content, for example, may be inauthentically amplified with paid or automated engagement, or by actors who are not what they seem. Similarly, entirely authentic actors—domestic political activists, perhaps—may use inauthentic automation. In discussing the potential impact of AI on future influence or disinformation campaigns, we therefore consider its potential for transforming each of the three factors. We believe that generative models

[75] François, *Actors, Behaviors, Content: A Disinformation ABC Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses*.

[76] François.

[77] Click farms refers to labor hired to manually click on content online on behalf of their employers. They display some online patterns of genuine internet users since they are humans, allowing them to avoid bot detection, while still driving up content views and interactions.

[78] Deepfake videos have already been used for phishing campaigns and the harassment of journalists. Some have suggested deepfakes may be used to develop crisis scenarios, whether by faking government directives, discrediting candidates for public office, or pretending to keep hostage soldiers. See, for example, Kishalaya Kundu, "Criminals Used AI To Clone Company Director's Voice And Steal $35 Million," Screen Rant, October 14, 2021, https://screenrant.com/ai-deepfake-cloned-voice-bank-scam-theft-millions/; Katerina Sedova et al., *AI and the Future of Disinformation Campaigns: Part 2: A Threat Model* (Center for Security and Emerging Technology, December 2021), https://doi.org/10.51593/2021CA011; Rana Ayyub, "I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me," *Huffington Post*, November 21, 2018, https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316; Jan Kallberg and Stephen Col. Hamilton, "US military must prepare for POW concerns in the deepfake era," C4ISRNET, August 23, 2021, https://www.c4isrnet.com/opinion/2021/08/23/us-military-must-prepare-for-pow-concerns-in-the-deepfake-era/.

will improve the content, reduce the cost, and increase the scale of campaigns; that they will introduce new forms of deception like tailored propaganda; and that they will widen the aperture for political actors who consider waging these campaigns. In Table 6, we summarize possible changes to the actors, behavior, and content due to language models, and describe these changes in further depth below.

ABC	Potential Change Due to Generative AI Text	Explanation of Change
Actors	Larger number and more diverse group of propagandists emerge.	As generative models drive down the cost of generating propaganda, more actors may find it attractive to wage influence operations.
	Outsourced firms become more important.	Propagandists-for-hire that automate production of text may gain new competitive advantages.
Behavior	Automating content production increases scale of campaigns.	Propaganda campaigns will become easier to scale when text generation is automated.
	Existing behaviors become more efficient.	Expensive tactics like cross-platform testing may become cheaper with language models.
	Novel tactics emerge.	Language models may enable dynamic, personalized, and real-time content generation like one-on-one chatbots.
Content	Messages grow more credible and persuasive.	Generative models may improve messaging compared to text written by propagandists who lack linguistic or cultural knowledge of their target.
	Propaganda is less discoverable.	Existing campaigns are frequently discovered due to their use of copy-and-pasted text (copypasta), but language models will allow the production of linguistically distinct messaging.

Table 6: How Language Models May Influence the ABCs of Influence Operations

4.1.1 Actors: Outsourced Execution & Proliferation of Propagandists

One limitation on actors who run disinformation campaigns is cost. While social media has decreased the cost to reach the public, most campaigns have involved numerous fake personas, sophisticated automation, and/or a stream of relevant content. AI reduces the cost of running campaigns further, by automating content production, reducing the overhead in persona creation, and generating culturally appropriate outputs that are less likely to carry noticeable markers of inauthenticity. These developments will expand the set of actors with the capacity to run influence operations.

The notion that less resourced actors (or less talented trolls) could use AI models to run influence operations is not merely speculative—it has already been piloted. Recently, a researcher fine-tuned a model hosted on HuggingFace (an online hub for machine learning models) on a dataset of 4chan posts[79] and

[79] Matt Murphy, "Someone trained an A.I. with 4chan. It could get worse.," Slate, August 3, 2022, https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html.

dubbed it “GPT-4chan.” He proceeded to post more than 30,000 generated posts on 4chan.[80] In this case, the original model was publicly available and easily downloadable. In another example, in October 2019, Idaho solicited public feedback about a proposal to change its Medicaid program. A Harvard Medical School student ran a study in which he submitted comments that were generated by GPT-2 as if they were written by ordinary citizens. In a follow-on survey, volunteers were unable to distinguish between the AI-generated and human-written comments.[81] If a single student can run this type of campaign on a public comment board, political actors will likely be able to do the same, leading to a wider pool of potential actors waging influence operations.[82]

Independently of improvements in generative AI models, political actors are increasingly turning toward third-party influence-for-hire companies to conduct their campaigns, including firms that otherwise appear to be legitimate marketing or PR firms.[83] Even if AI companies place restrictions on who can access their models, this trend makes it harder to ensure that bad actors do not have access to generative models, as marketing firms will likely be granted access given their other legitimate uses.[84]

4.1.2 Behavior: Low-Cost Content at Scale and Novel Techniques

In addition to affecting the actors involved in influence operations, the integration of generative language models can encourage new types of behaviors used in influence campaigns and change the way existing behaviors are enacted in practice.

The most basic behavioral change that will result from using language models for influence operations is replacing, or augmenting, a human writer in the content generation process. Language models replacing human writers, or used in a human-machine team, could dramatically reduce the cost and increase the scalability of the types of propaganda campaigns we see today—such as mass-messaging campaigns on social media platforms or long-form news generation on unattributable websites.

Beyond simply writing text, generative models can improve other existing tactics, techniques, and procedures of influence operations. For instance, cross-platform testing is a long-standing component of many influence operations, in which actors first test content on one platform to gauge audience reaction before proliferating content onto other platforms.[85] Operators using generative AI models may be able to perform this type of testing at greater scale, which may improve a campaign’s overall impact.

Manipulative actors could also use language models to overwhelm or falsify checks in areas in which text commentary is solicited, such as in the public comment process between governments and their

[80] Andrey Kurenkov, “Lessons from the GPT-4Chan Controversy,” The Gradient, June 12, 2022, https://thegradient.pub/gpt-4chan-lessons/; James Vincent, “YouTuber trains AI bot on 4chan’s pile o’ bile with entirely predictable results,” *The Verge*, June 8, 2022, https://www.theverge.com/2022/6/8/23159465/youtuber-ai-bot-pol-gpt-4chan-yannic-kilcher-ethics.

[81] Will Knight, “AI-Powered Text From This Program Could Fool the Government,” Wired, January 15, 2021, https://www.wired.com/story/ai-powered-text-program-could-fool-government/.

[82] As we discussed in Section 2, GPT-2 is already publicly available, as are stronger models like Eleuther’s GPT-NeoX-20B, a 20-billion parameter model.

[83] See: Josh A. Goldstein and Shelby Grossman, “How disinformation evolved in 2020,” January 4, 2021, https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/; Max Fisher, “Disinformation for Hire, a Shadow Industry, Is Quietly Booming,” *New York Times*, July 25, 2021, https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html.

[84] Sedova et al., *AI and the Future of Disinformation Campaigns: Part 2: A Threat Model*.

[85] *Senate Report No 116-290, vol 2* (2020), https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf.

citizens.[86] Recent research showed that public comments to the Federal Communications Commission about net neutrality in 2017 were largely driven by falsified repeated comments.[87] Language models may increase the scale and decrease detectability of similar future operations. In a recent field experiment, researchers sent over 30,000 emails—half written by GPT-3, and half written by students—to 7,132 state legislators. The researchers found that on some topics legislators responded to computer-generated content at only a slightly lower rate than human-generated content; on other topics, the response rates were indistinguishable.[88]

Language models will also shape propagandists' behaviors by introducing new behaviors altogether and enabling novel tactics. Because these models make it possible to "think up" a new version of content in near real time, actors can deploy them for real-time, dynamic content generation. In the next few years, as language models improve, it may be possible for propagandists to leverage demographic information to generate more persuasive articles that are strongly tailored to the target audience.

Whether this will be a cost-effective strategy is dependent on how well models (or future models) can tailor messaging based on limited demographic information. Today, websites could use demographic information to route users to different human-written articles. Writing different versions of articles, however, takes human capital. Language models, by contrast, could provide original articles for each combination of user demographics, which would be infeasible for human writers. The payoff of this strategy depends on how persuasive AI-generated text is, and how much more persuasive highly tailored personalized text is, compared to one (or a few) human-written articles. It could also involve humans making minor adjustments to AI-generated text. This remains uncertain but warrants further attention, as analogous personalization could be applied to a range of malicious campaigns, including phishing emails.[89]

Another central example of dynamic content generation is chat—language models engaging in extended back-and-forth conversations. Actors could potentially deploy personalized chatbots that interact with targets one-on-one and attempt to persuade them of the campaign's message.[90] This capability could materialize as interactive social media personas, back-and-forth email messaging, or faked support chatbots. Propagandists may leverage chat with language models across a wide range of contexts—anywhere interactivity is useful.

There are reasons to think that chat may be an important vector of influence. Researchers have already found that interacting with a chatbot can influence people's intentions to get a COVID-19 vaccine;[91] with chatbots based on language models, these interactions could be even more powerful. While deploying their own chatbots would give influence operators more control, they may be able to manipulate innocu-

[86] Knight, "AI-Powered Text From This Program Could Fool the Government."

[87] "Public Comments to the Federal Communications Commission about Net Neutrality Contain Many Inaccuracies and Duplicates," *Pew Research Center*, November 29, 2017, https://www.pewresearch.org/internet/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

[88] Sarah Kreps and Doug Kriner, "The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment," *(Working Paper)*.

[89] Andrew J. Lohn and Krystal A. Jackson, *Will AI Make Cyber Swords or Shields?* (Center for Security and Emerging Technology, August 2022), https://doi.org/10.51593/2022CA002.

[90] For a rudimentary example of a chat application built on language models, see "Marv the Sarcastic Chat Bot," OpenAI API, https://beta.openai.com/examples/default-marv-sarcastic-chat

[91] Sacha Altay et al., "Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions.," *Journal of Experimental Psychology: Applied*, October 28, 2021, ISSN: 1939-2192, https://doi.org/10.1037/XAP0000400.

ous chatbots to spread propaganda. Microsoft's Tay is one historical example,[92] and more sophisticated techniques to "poison" language models are being investigated by researchers.[93]

4.1.3 Content: High Quality and Low Detectability

There are two varieties of textual content commonly observed in influence operations: short-form commentary such as tweets or comments, and long-form text. Language models could improve the quality and therefore decrease the detectability of both types of content.

Short-form content is primarily pushed out by inauthentic account personas on social media, or sometimes in the comment sections of websites or blogs, and is often intended to influence the reader's perception of public opinion. Many tweets or comments in aggregate, particularly if grouped by something like a trending hashtag, can create the impression that many people feel a certain way about a particular issue or event. Producing this content, which purports to represent the opinions of the "man-on-the-street," requires account operators to have knowledge of the communication style and rhetoric that fits the persona who is purportedly speaking; some operations are exposed because of incongruities or "uncanny valley" dynamics in which the persona uses terminology or slang that does not quite fit what a genuine member of the community would likely say.[94]

Creating the appearance of a public opinion requires having many speakers. In 2014–2016, political operatives frequently used bots—automated accounts—to produce this volume, deploying them to make content trend or to introduce particular opinions into hashtags.[95] However, creating speech for large networks of automated accounts was a challenge, and the bot networks were often detectable because they used "copypasta"—repetitive or identical language across networks and accounts. In response, Twitter changed the weighting function for its trending algorithm to minimize the effect of bot accounts.[96] Subsequent takedowns suggest that some well-resourced state propagandists have shifted away from automated account networks posting copypasta or attempting to flood hashtags and toward more well-developed, non-automated persona identities.[97] Others did continue to leverage bots, though often to create the perception of engagement slightly differently, such as by replying to, retweeting, or liking tweets.

[92] Oscar Schwartz, "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation," *IEEE Spectrum*, November 25, 2019, https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[93] Eugene Bagdasaryan and Vitaly Shmatikov, "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures," *2022 IEEE Symposium on Security and Privacy*, 2022, 769–786, https://doi.org/10.1109/SP46214.2022.9833572.

[94] On the idea of an uncanny valley, see Tom Geller, "Overcoming the Uncanny Valley," *IEEE Computer Graphics and Applications* 28, no. 4 (July-Aug. 2008): 11–17, ISSN: 02721716, https://doi.org/10.1109/MCG.2008.79. For evidence that technology has surpassed the uncanny valley for producing as-if human faces, see Sophie J. Nightingale and Hany Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *PNAS* 119, no. 8 (February 2022), ISSN: 10916490, https://doi.org/10.1073/PNAS.2120481119

[95] Samuel C. Woolley and Douglas Guilbeault, "Computational propaganda in the United States of America: Manufacturing consensus online," *Project on Computational Propaganda Research*, 2017, 1–29.

[96] Ed Ho, "An Update on Safety," Twitter Blogs, February 7, 2021, https://blog.twitter.com/en_us/topics/product/2017/an-update-on-safety.

[97] Renee DiResta et al., "In Bed with Embeds: How a Network Tied to IRA Operations Created Fake "Man on the Street" Content Embedded in News Articles," *Stanford Internet Observatory*, December 2, 2021, https://cyber.fsi.stanford.edu/io/publication/bed-embeds; Shelby Grossman, Khadija H., and Emily Ross, *Royal Sockpuppets and Handle Switching: How a Saudi Arabia-Linked Twitter Network Stoked Rumors of a Coup in Qatar* (Stanford Internet Observatory, October 2020), https://stacks.stanford.edu/file/druid:hp643wc2962/twitter-SA-202009.pdf.

As generative AI models continue to advance, they could make it possible for influence operators to automate the generation of text commentary content that is as varied, personalized, and elaborate as human-generated content. If propagandists can use generative models to produce semantically distinct, narratively aligned content, they can mask some of the telltale signs (identical, repeated messaging) that bot detection systems rely on—prompting bot detection systems to leverage other signals. This evolution could allow even small groups to make themselves look much larger online than they are in real life.

Real IRA Tweet	Generated Tweet
Shocking Video 😮😮😮 US police repeatedly tasing a black man holding his baby in his own apartment in Phoenix, Arizona. We're not safe in this country. We're nor safe in our own homes! #BlackLivesMatter #PoliceBrutality #Police https://t.co/ldWNFWOADg	This video is everything that's wrong with the police. They act like a pack of wolves, trying to scare this man away. It's unacceptable! https://t.co/ldWNFWOADg

Table 7: For short-form text, large language models can already match the capabilities of human-written segments in real influence operations. The left tweet is the top-performing tweet by number of retweets in an IRA-backed Ghanian disinformation campaign released by Twitter in March 2020. The right tweet is generated by prompting a language model with a few example tweets and then asking it to produce a tweet with the given link.

A second relevant output of language models for influence operations is long-form text, such as propagandistic journalism. This content is used to make a longer point, and often appears on front media properties, such as gray media outlets owned or controlled by the disinformation actor or undisclosed allies. Often, one of the goals is to have the claims in the text republished by more reputable authentic sources, a technique known as "narrative laundering." For example, Russia's "Inside Syria Media Center" (ISMC) news website, a GRU front property whose bylined journalists included fabricated personas, produced content that was republished as contributed content within ideologically aligned, unwitting publications, or incorporated into real news articles in the context of expert quotes.[98]

Producing this kind of long-form propaganda, however, takes time and expertise. The inauthenticity of the ISMC was uncovered when the GRU's inauthentic journalist personas began to plagiarize each other's work; an editor from one of the publications that received a submission from an ISMC journalist inquired about the apparent plagiarism, then began to investigate the site after receiving an incongruous response. Learning from this experience, threat actors affiliated with the Russian IRA reverted to old-school methods and hired unwitting freelance journalists to write for proxy outlets; they, too, were uncovered when the journalists began to look more deeply into the publications.[99] Language models, however, can produce long-form content in seconds, reducing the time, cognitive load, and cost to produce such content and eliminating the need to take risky shortcuts—or hire real people—that might jeopardize the overall operation. The novel behavior—deployment of generative models—improves the

[98] Renée DiResta, Shelby Grossman, and Alexandra Siegel, "In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies," *Political Communication* 39, no. 2 (2021): 222–253, ISSN: 10917675, https://doi.org/10.1080/10584609.2021.1994065.

[99] Jack Delaney, "I'm a freelance writer. A Russian media operation targeted and used me," *The Guardian*, September 4, 2020, https://www.theguardian.com/technology/2020/sep/04/russia-media-disinformation-fake-news-peacedata; *August 2020 Coordinated Inauthentic Behavior Report* (Meta, September 1, 2020), https://about.fb.com/news/2020/09/august-2020-cib-report/; Jack Stubbs, "Russian operation masqueraded as right-wing news site to target U.S. voters," Reuters, October 1, 2020, https://www.reuters.com/article/usa-election-russia-disinformation/exclusive-russian-operation-masqueraded-as-right-wing-news-site-to-target-u-s-voters-sources-idUSKBN26M5OP.

quality of long-form text that could increase the impact of these campaigns.

There is already some evidence that existing language models could substitute for human authors in generating long-form content or make content generation more effective through human-machine teaming. In a series of survey experiments, researchers found that GPT-2, the smaller predecessor of GPT-3, could produce text that successfully mimicked the style and substance of human-written articles.[100] In experiments of GPT-3's capabilities, human participants were able to distinguish multiparagraph GPT-3 "news articles" from authentic news articles at a rate only slightly better than random chance.[101] In an experimental setting, researchers also found that GPT-3-generated propaganda articles were nearly as persuasive as articles from real world covert propaganda campaigns.[102] Language models could also be used to generate summary text of other articles, inflected for ideological alignments.

It seems likely that language models are cost-effective (relative to human propagandists) for some campaigns. For a simple calculation to demonstrate this claim, let w represent the hourly wage paid to information operators, L_h represent the productivity of human authors (measured as the number of posts that can be written by a human in an hour), c represent the amortized per-output cost of generating posts using a language model, and L_r represent the productivity of human reviewers (measured as the number of AI-generated posts that a human can review in an hour). Further, let p represent the proportion of AI outputs that are "usable" for an information operation. Then, the cost of generating n outputs will be equal to $\frac{n*w}{L_h}$ in the case of a human campaign, and $(c + \frac{w}{L_r}) * \frac{n}{p}$ in the case of an AI-augmented campaign where humans are tasked to read and approve AI outputs.

The amortized per-output cost of producing content may be relatively high in cases where a large language model is trained from scratch and used for a short campaign, but if a public model is used or a model is trained and reused for sufficiently many campaigns, c will approach the bare electricity cost of operating the model, which can be negligible compared to the human labor costs of either authoring or reviewing outputs. In this case, the AI-augmented campaign will be more cost effective than a fully human one, so long as the inequality

$$L_r/L_h > 1/p$$

holds. In other words, so long as the ratio between the number of posts that a human can review in an hour and the number of posts that a human can write in an hour is larger than the number of AI-generated posts that a human must review, on average, to get one usable output, then the use of the AI model will be cost-effective. Only very moderate assumptions are needed to make this inequality hold; for example, if outputs from current language models are passably coherent and usable for some (possibly unsophisticated) operations more than 20% of the time, then this inequality will hold as long as a human could read at least five posts in the time it takes to author one.[103]

[100] Kreps, McCain, and Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation."

[101] Tom B. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (May 2020), ISSN: 10495258, https://doi.org/10.48550/arxiv.2005.14165.

[102] Goldstein et al., "Can AI write persuasive propaganda?"

[103] For a more extended analysis of this topic, see Musser, "A Cost Analysis of Generative Language Models and Influence Operations"

4.2 Expected Developments and Critical Unknowns

Both the recent technological progress in generative models and their wider diffusion are likely to continue. Here we speculate on several expected technological developments over the coming years that will be major drivers of operational change. We also highlight critical unknowns, where multiple paths are possible, and where this uncertainty may have a large impact on the future state of play. These projections are not intended as explicit forecasts, but rather as a way to conceptualize medium-term plausible futures. This section is summarized in Table 8.

Technical and Strategic Unknowns	**Current State (2022)**	**How This Might Change**
Usability, reliability, and efficiency of generative models	•Difficult to specify and stay on a task •Outputs can be incoherent or fabricate facts •Building models from scratch can cost millions of dollars; efficacy of fine-tuning still being explored for different capabilities	•Train to better follow instructions •Retrain periodically on fresher data •Hardware, software, and engineering progress
Difficulty of developing new and more general capabilities relevant to influence operations	•Can produce tweets, short news articles •Little interactivity or long-range dialogue •Not optimized for influence (via proxies like click-through rate)	•Scaling up with bigger models and more data •Using metrics of influence to train models •Combining models with non-ML software pipelines and human reviewers
Interest and investment in AI for influence; accessibility of text generation tools	•Leading AI R&D mostly done by industry labs and academic institutions in a few countries for scientific or commercial merit •No free online tools to generate arbitrary state-of-the-art text at scale	•Nation-state invests in or adapts AI for influence •Marketing industry adopts language models •State-of-the-art language model published online with an easy user interface, free for anyone to use

Table 8: Expected Developments For Generative Models In Influence Operations

4.2.1 Improvements in Usability, Reliability, and Efficiency

Language models are likely to improve on three features that will affect their deployment in influence operations: **usability** (how difficult it is to apply models to a task), **reliability** (whether models produce outputs without obvious errors), and **efficiency** (the cost-effectiveness of applying a language model for influence operations).

Improvements in usability and reliability could allow lower-skilled propagandists to employ language models with reduced human oversight. Achieving existing capabilities—like writing slanted short articles or tweets—will become much cheaper and more efficient, which could increase the rate of adoption of language models in influence operations.

Usability

While recent generative models have become more generalizable—users can specify a wide range of tasks—it takes skill and experience for the user to operate the model successfully. For example, it is difficult for an operator to specify a task for a language model. Imagine prompting a language model with the input "What is 15 times 37?" To an operator, it may be obvious that the output for this prompt should be a single number (555), but to the model—which by default is simply performing a text completion task—an equally plausible continuation of this text may be "What is 89 times 5?" as though the task it had been assigned was to write a list of exam questions for a grade school math course. Prompt engineering, where operators experiment with different ways of phrasing their requests, can help mitigate this problem, but it can only go so far without the ability to fine-tune or otherwise alter the base model itself.[104]

Researchers are exploring different approaches to improve task specification. For example, some researchers have modified the training process of large language models to improve the ability of those models to follow instructions.[105] Other researchers have tried tagging different parts of the training data by their types (e.g., "dialogue" would specify dialogue data), and then asking a model to only produce data of a certain type.[106] Other approaches are in development,[107] and it remains unclear which combination of approaches will ultimately be adopted. If usability of language models improves, propagandists will be able to use models for new tasks as they arise without in-depth prompt engineering experience. Furthermore, because it is often difficult to predict which tasks a language model can be used for, improvements in usability can make it easier for propagandists to experiment with and discover applications of language models in influence operations.

Reliability

Language models can generate plausible content for a wide variety of tasks. However, even if plausible content is initially generated, a propagandist must either trust that a model will be highly reliable—completing the task without making detectable errors—or apply consistent monitoring. But mod-

[104] See Pengfei Liu et al., "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, September 2021, https://doi.org/10.1145/3560815, and "Prompt Engineering," co:here, https://docs.cohere.ai/docs/prompt-engineering for a popular explanation.

[105] Long Ouyang et al., "Training language models to follow instructions with human feedback," *OpenAI*, March 2022, https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

[106] Nitish Shirish Keskar et al., "CTRL: A Conditional Transformer Language Model for Controllable Generation," *arxiv:1909.05858 [cs.CL]*, September 2019, https://doi.org/10.48550/arxiv.1909.05858.

[107] For a broad overview of some approaches to this problem, see: Lilian Weng, "Controllable Neural Text Generation," Lil'Log, January 2, 2021, https://lilianweng.github.io/posts/2021-01-02-controllable-text-generation/.

els are often not reliable, and consistent monitoring introduces additional costs. As task complexity increases,[108] ensuring compliance becomes increasingly difficult. If models fail to consistently produce compelling outputs, propagandists may simply choose not to use them. These challenges then increase the demand for more skilled operators, who may be in short supply. An important caveat, however, is that not every task may require the same level of reliability. For example, deploying Twitter bots that sometimes produce incoherent tweets might be fine for a propagandist if the goal is to simply cause chaos around a targeted topic.[109]

Unreliable outputs show up in different forms, but the core takeaway is that although language models can produce high-quality multiple-page documents, they cannot do so consistently. Common failure modes include extremely repetitive outputs, losing coherency over the course of a long output, or fabricating stylized facts that do not fit the generation context.[110]

One reason why models fail to consistently produce high-quality text is because they lack awareness of time and information about contemporary events. The current training regime for generative models trains them once on a large corpus of data, which means that models will not have context for events that occur after this key moment.[111] Ask a language system that was trained before COVID-19 about COVID-19, and it will simply make up plausible-sounding answers without any real knowledge about the events that unfolded.

To address the problem of a lack of up-to-date information, AI researchers will likely pursue two basic approaches: either continually retrain models to account for new context, or develop new algorithms that allow for more targeted updates to a language model's understanding of the world.[112] For instance, language models that are trained to be "time aware" can perform much better at handling recent trends, references to named entities, and concept drift—the way in which words can change in meaning overtime.[113] Since propagandists may be interested in shaping the perception of breaking news stories, significant improvements in how language models handle recent events not present in their initial training data will translate directly into improved capabilities for influence operators across a wide number of potential goals.

State-backed propagandists will also likely be interested in methods to adapt pretrained language models to new tasks, which would give them some assurance of reliability. Current methods to adapt models to new tasks require examples of those tasks, and use the examples to fine-tune a model to handle them well. For example, if a model performs unreliably on Spanish-language inputs, one might fine-tune that model on more examples of Spanish text.

Efficiency

Alongside improvements to usability and reliability, we expect improvements in the efficiency of language models, which will reduce the costs to automate some influence tactics. Models that can more

[108] For example, imagine trying to convey to a model that its task is to take headlines and subtly rewrite them to be consistently biased toward a certain political ideology.

[109] And if these errors do not make it easier to attribute or detect inauthentic behavior.

[110] Ari Holtzman et al., "The Curious Case of Neural Text Degeneration," *arxiv:1904.09751 [cs.CL]*, February 19, 2019, ISSN: 16130073, https://doi.org/10.48550/arxiv.1904.09751.

[111] Bhuwan Dhingra et al., "Time-Aware Language Models as Temporal Knowledge Bases," *Transactions of the Association for Computational Linguistics* 10 (March 2022): 257–273, ISSN: 2307387X, https://doi.org/10.1162/tacl_a_00459.

[112] One example of this is what are known as retrieval-based methods, in which a language model is trained to retrieve knowledge from an external database. To achieve time-awareness, operators may simply need to update that external database.

[113] Daniel Loureiro et al., "TimeLMs: Diachronic Language Models from Twitter," *arxiv.2202.03829 [cs.CL]*, February 2022, 251–260, https://doi.org/10.48550/arxiv.2202.03829.

efficiently guess the next word for marketing copy can also more efficiently guess the next word for a polarizing article. Efficiency gains could come from many angles: algorithmic progress, hardware improvements, or the use of inexpensive fine-tuning to optimize relatively small models for influence operation-specific tasks.[114]

Other future improvements in the influence operations space could include organizational and operational innovation. Organizations may improve human-machine collaboration by creating software that improves a propagandist's ability to oversee, select, and correct the outputs of language models. Language models could be used as an autocorrect for cultural context, allowing operators to work with targets they are not familiar with, and allowing familiar actors to output a higher volume of credible content per unit time.

The empirical details of efficiency will be important. Exactly how efficiently can generative models be trained? One measure of algorithmic progress in image classification found a 44x improvement over the course of nine years.[115] Even during the course of drafting this paper, research has come out that claims to train GPT-3 quality models for less than $500,000, which would represent a factor of 3–10x improvement.[116] If capabilities relevant to influence operations—generating persuasive text, fake personas, or altered videos—are achievable with significantly lower cost, then they are more likely to diffuse rapidly. Similarly, how efficient will an operation be by using language models as a complement to human content editors, rather than as a full substitute? The operational know-how and ease of editing might make it easier to scale up influence operations.

4.2.2 New and More General Capabilities for Influence

As language models improve, it is likely that they will have newer and more general capabilities. In 2017, few expected that language models in 2022 would be able to add and multiply three-digit numbers without having been trained to do so.[117] Not surprisingly, we do not know what capabilities the language models of 2027 will have.

In this section we discuss two critical unknowns related to this theme:

1. Which capabilities will emerge as side effects of scaling to larger models? If abilities directly applicable to influence operations—such as the ability to persuade via long-lasting dialogue—emerge as a side effect of simply scaling to larger models, then many AI projects are high risk—regardless of the goals of their creators.

2. How difficult is it to train generative models to execute the various capabilities that are useful for influence operations? If it is easy for generative models to learn skills (like writing viral or

[114] On fine-tuning GPT-2, a smaller language model, to mimic several news sources with high accuracy, see Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation* 14-15. Recent research has also explored more efficient methods of fine-tuning models, which could make it even easier to fine-tune models for influence operations tasks.

[115] By one measure, between 2012 and 2019, algorithmic efficiency doubled every 16 months on average. The number of floating-point operations required to train a classifier to a given level decreased by a factor of 44x; see Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," *arxiv:2005.04305 [cs.LG]*, May 2020, https://doi.org/10.48550/arxiv.2005.04305

[116] Venigalla and Li, "Mosaic LLMs (Part 2): GPT-3 quality for <$500k."

[117] Jason Wei et al., "Emergent Abilities of Large Language Models," *arxiv:2206.07682 [cs.CL]*, June 2022, https://doi.org/10.48550/arxiv.2206.07682.

persuasive text) for influence operations, then the problem of defense becomes more urgent.

New Capabilities as a Byproduct of Scaling and Research

New capabilities for influence operations may emerge unexpectedly as language models are scaled up. One of the impressive scientific takeaways from recent progress in generative models is that training on a simple objective—predicting the next word or pixel—gives rise to adjacent, general capabilities. A system trained to predict the next word of an input text can also be used to summarize passages or generate tweets in a particular style; a system trained to generate images from captions can be adapted to fill in parts of a deleted image, and so on. Some of these abilities only emerge when generative models are scaled to a sufficient size.[118]

Today, we have single language systems that can summarize short texts, translate between languages, solve basic analogies, and carry on basic conversations; these capabilities emerged with sufficiently large language models.[119] It is difficult to predict when new capabilities will emerge with more scaling or even whether a given capability is present in a current system. Indeed, in a salient recent example, an engineer from Google became persuaded that the Google model he was interacting with was sentient.[120] These sorts of emergent capabilities seem hard to anticipate with generative models, and could be adapted by influence operators.

Even more generally, as more actors begin to work on AI development with different motivations and in different domains, there is a possibility that some capabilities emerge as side effects of research. Because much AI development attempts to target more general capabilities, a small adjustment might suffice to uncover capabilities relevant to influence operations. For example, improvements in reasoning capabilities might also allow generative models to produce more persuasive arguments.

Models Specialized for Influence

Above, we described the possibility that scaling will (unintentionally) make language models better tools for influence operations. Another possibility is that propagandists will intentionally modify models to be more useful for tasks like persuasion and social engineering. Here, we mention three possible paths of improvement: targeted training, generality, and combinations with other technologies.

The first likely improvement is targeted training. Generative models could be trained specifically for capabilities that are useful for influence operations. To develop these capabilities, perpetrators may choose to incorporate signals such as click-through data or other proxies for influence. These signals may be included in the training process, resulting in generative models more strongly optimized to produce persuasive text. Advertising and marketing firms have economic incentives to train models with this type of data, and may inadvertently provide the know-how for propagandists to do the same. Another form of targeted training would be to withhold or modify the information in the training data to affect how the trained model produces content. For example, suppose that a language model is trained with all mentions of a particular group occurring alongside false negative news stories. Then even innocuous deployments of products based on that language model–like a summarizer or customer support chatbot–may produce slanted text without being transparent to model users.

[118] Wei et al., "Emergent Abilities of Large Language Models."

[119] Ganguli et al., "Predictability and Surprise in Large Generative Models."

[120] Nitasha Tiku, "The Google engineer who thinks the company's AI has come to life," *Washington Post*, June 11, 2022, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Targeted training may be less resource-intensive than training more general models. The difficulty of automating specific tasks is challenging to estimate and often defies intuition.[121] There is some preliminary evidence already that systems like GPT-3 can write slanted news articles—without being explicitly trained for that task.[122] It may be possible for future systems to be engineered to write extremely persuasive, tailored texts, or carry on long-lived dialogue.

In addition to targeted training, improvements in the generality of model capabilities are likely to have applications to influence operations. For example, one improvement in generality comes from simply combining different modalities into a single system: a single model that can consume and generate both images and text, for example. One can imagine instructing a bot built on such a system to ingest images on the internet, cleverly respond to them, produce completely fabricated images, and carry on a conversation—all at the same time.

Finally, a more prosaic path to achieving new capabilities would be to simply combine generative models with other forms of automation. It is possible that using generative models as the "engine" for intelligent bots, along with software to accommodate for shortcomings, could lead to more human-like behavior. For example, a propagandist could write software to find and copy the Facebook profiles of people with interests compatible with the propaganda message, and use this to prompt the generative model. The development of this system may also benefit from integrating software that has already been developed separately, perhaps by chaining together smaller language models.[123]

4.2.3 Wider Access to AI Capabilities

In understanding the impact of language models on influence operations in the future, a key consideration is which actors will have access to language models and what may precipitate their use in influence operations. We highlight three critical unknowns in this domain:

1. Willingness to invest in state-of-the-art generative models. Right now, a small number of firms or governments possess top-tier language models, which are limited in the tasks they can perform reliably and in the languages they output. If more actors invest in state-of-the-art generative models, then this could increase the odds that propagandists gain access to them. It is also possible that uncertain and risky investments could lead to the creation of systems that are much better at tasks relevant to influence operations.

2. The existence of unregulated tooling. Proliferation of easy-to-use interfaces to generate persuasive text or images can increase the adoption of generative models in influence operations. If these tools are developed, we are likely to see an earlier and broader uptick of generated content in influence operations.

3. Intent-to-use generative models for influence operations. As access to generative models increases, an actor's willingness to use these models in influence operations might be an important constraint.

[121] This observation is related to the well-known Moravec's paradox: "Moravec's paradox," Wikipedia, accessed June 29, 2022, https://en.wikipedia.org/wiki/Moravec%5C%27s_paradox.

[122] For example, in some experiments to produce slanted text with GPT-3 in 2021, researchers experimented with generating articles from sources such as *The Epoch Times*; see Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation*.

[123] Tongshuang Wu et al., "PromptChainer: Chaining Large Language Model Prompts through Visual Programming," *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, April 2022, https://doi.org/10.1145/3491101.3519729.

> If social norms do not constrain the use of models to mislead, then actors may be more likely to deploy models for influence operations.

Willingness to Invest in Generative Models

In Section 4.2.2, we outlined ways that language models could be leveraged for influence operations. First, propagandists could repurpose (or steal) state-of-the-art models with new and more general capabilities. Second, sophisticated propagandists could train models specifically for influence operations. In both cases, the application of generative models to influence operations may ultimately be constrained by different actors' willingness to make large and potentially risky investments in developing generative models.

To have an impact on influence operations, a large investment need not target generative models for influence operations specifically. An investment could simply target more general generative models for other purposes such as scientific discovery or commercial value. If many actors—such as governments, private firms, and even hyperwealthy individuals—develop these state-of-the-art language models, then that increases the odds that propagandists could gain access (legitimately or via theft) to models that can be repurposed for influence operations. For example, a propagandist could fine-tune a stolen model to produce persuasive text in different languages or in a particular domain.

In the extreme case, the propagandist themself could be a well-resourced actor—like a determined country—and make a risky and large investment in developing a generative model-based system specifically for influence operations. This may require extensive computational resources, bespoke data—such as user engagement metrics—and engineering talent. In either case, it may not be clear how feasible some engineering projects are; the timeline for advances may ultimately depend on whether propagandists decide to make uncertain investments in developing these generative models.

While there are reasons why well-resourced actors might make large investments in developing models for influence, there are also reasons to forgo them. We are already reaching the point where the creation of convincing tweet-sized texts can be automated by machines. However, there could be diminishing returns for influence operations for more advanced capabilities, which would make large investments by propagandists specifically unlikely. For example, if most influence operations rely on a deluge of similarly short bits of content to sway attention-bound humans, there may be few incentives to develop generative models that can generate longer pages of human-like text.

Greater Accessibility from Unregulated Tooling

Even with nominal access to models, there will likely be some operational know-how required to use them. For example, applying GPT-3 to propaganda tasks requires fiddling with the exact inputs you give the system. To create a photorealistic image a few years ago, a propagandist would have had to run a model themselves on their own infrastructure. But packaging easy-to-use tools that do these tasks has since lowered the operational know-how required to apply generative models to influence operations. Today, anyone with access to the internet can obtain photorealistic AI-generated images from websites such as thispersondoesnotexist.com. AI-generated profile pictures (images of people) are now commonplace in influence operations[124] and have also been used for deceptive commercial purposes.[125] It is quite possible that had this easy-to-use tooling not been developed, influence operations would not have leveraged AI-generated profile pictures to add plausibility to their campaigns, or may not have done so to the same extent.

An analogous lesson may apply to the use of language models for influence operations as well. If easy-to-use tools for language models proliferate, we may see propaganda campaigns rely on language models (that would otherwise not have). Easy-to-use tools that produce tweet- or paragraph-length text could lower the barrier for existing propagandists who lack machine learning know-how to rely on language models. Easy-to-use tools could also lead to the integration of new capabilities, such as automated chatbots deployed to troll targets determined by a bad actor. At the same time, the creation of easy-to-use language model tools could also lead to the proliferation of propagandists. Firms and private individuals who may once have avoided waging propaganda campaigns could now choose to do so because of declining costs.

Norms and Intent-to-use

The intent (or lack thereof) may be an important constraint on the application of generative models to influence operations. In the political science literature, a norm is a "standard of appropriate behavior for actors with a given identity."[126] Scholars describe three stages for a norm to take hold internationally: norm emergence (a norm is built by norm entrepreneurs, or "people interested in changing social norms"[127]), a norm cascade (more countries rapidly adopt the norm), and internationalization of the norm (a norm becomes widely accepted and taken for granted.[128]) Studies show that norms constrain different types of state behavior that would be expected to take place by a cost-benefit analysis. International security scholars have argued that norms have powerfully restrained state behavior—from using nuclear weapons, from more routine use of assassinations, and from widespread use of mercenaries.[129]

The notion that norms can constrain behavior in different facets of domestic and international life may provide a useful lesson for the use of language models for influence operations. Even if an actor has access to models that can easily be repurposed to create persuasive chatbots, and even if this can be

[124] Shannon Bond, "AI-generated fake faces have become a hallmark of online influence operations," *NPR*, December 15, 2022, https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations.

[125] Josh A. Goldstein and Renée DiResta, "This salesperson does not exist: How tactics from political influence operations on social media are deployed for commercial lead generation," *Harvard Kennedy School Misinformation Review 3*, no. 5 (September 2022), https://doi.org/10.37016/MR-2020-104.

[126] Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change.," *International Organization* 52, no. 4 (1998): 887–917, https://www.jstor.org/stable/2601361. Norms involve two components: a prescription (what to do, or what not to do) and parameters (the situations under which the norm applies). For a description of this literature, see Vaughn P. Shannon, "Norms Are What States Make of Them: The Political Psychology of Norm Violation," *International Studies Quarterly* 44, no. 2 (June 2000): 293–316, ISSN: 0020-8833, https://doi.org/10.1111/0020-8833.00159.

[127] Cass R. Sunstein, "Social Norms and Social Roles," *Columbia Law Review* 44 (1996): 909, https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=12456&context=journal_articles.

[128] Finnemore and Sikkink, "International Norm Dynamics and Political Change."

[129] Tannenwald famously argued that non-use of nuclear weapons since the bombing of Hiroshima and Nagasaki cannot be explained by deterrence, but rather is the result of a normative prohibition on the use of nuclear weapons. See: Nina Tannenwald, "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use," *International Organization* 53, no. 3 (1999): 433–468, https://www.jstor.org/stable/2601286. (For evidence that challanges this theory, see Janina Dill, Scott D. Sagan, and Benjamin A. Valentino, "Kettles of Hawks: Public Opinion on the Nuclear Taboo and Noncombatant Immunity in the United States, United Kingdom, France, and Israel," *Security Studies* 31, no. 1 (2022): 1–31, ISSN: 15561852, https://doi.org/10.1080/09636412.2022.2038663; Sarah Percy, *Mercenaries: The History of a Norm in International Relations* (Oxford University Press, October 2007), 1–280, ISBN: 9780191706608

done at minimal cost to them, an actor must still decide to actually build and deploy them. Norms could constrain political actors from using language models for influence operations, and they could encourage developers to inhibit the use of language models for influence operations where possible.

Creating a norm that it is unacceptable to use language models for influence operations will likely require "norm entrepreneurs" to advocate this position. On the international level, this could be a coalition of states creating an agreement that they will not use language models for propaganda purposes. These states could devise mechanisms to punish those who fail to comply with the norm, or to reward those that join the coalition. On a substate level, machine language researchers or ethicists could also create a coalition to develop norms prohibiting the use of language models for influence operations. In fact, several AI researchers penned an open letter condemning activities like GPT-4chan,[130] explicitly citing the lack of community norms around the responsible development and deployment of AI as the reason to speak out.[131] Likewise, the marketing and PR industries could develop a norm against providing politicians AI-enabled influence operations as a service.

[130] We discussed this incident in Section 4.1.1. In brief, a researcher fine-tuned a publicly accessible language model on 4chan posts and proceeded to automatically post over 30,000 times in three days.

[131] Percy Liang, Rob Reich, and et al, "Condemning the deployment of GPT-4chan," accessed July 22, 2022, https://docs.google.com/forms/d/e/1FAIpQLSdh3Pgh0sGrYtRihBu-GPN7FSQoODBLvF7dVAFLZk2iuMgoLw/viewform?fbzx=1650213417672418119.

5 Mitigations

5.1 A Framework for Evaluating Mitigations

In this section, we move from describing the threat and attempt to outline a series of possible mitigations that could reduce the dangers of AI-enabled influence operations. Our goal here is to present a range of possible mitigations that various stakeholders could take to reduce the threat of AI-powered influence operations. Importantly, these mitigations are meant to be scoped to language models specifically, and we do not aim to articulate all the mitigations that could be taken to reduce the threat of misinformation generally.[132] Nevertheless, it is important to emphasize that, while generative models could help propagandists produce some types of harmful content, influence operations do not need AI models in order to succeed. As such, mitigations discussed here should be viewed as complements to broader and ongoing counter-influence operations efforts.

We group our mitigations based on four "stages" of the influence operation pipeline where they could be targeted: (1) model construction, (2) model access, (3) content dissemination, and (4) belief formation.[133] This grouping reflects that propagandists need four things to successfully use generative language models to shape the information ecosystem: first, there must be AI models capable of generating scalable and realistic-looking text; second, operators must have regular and reliable access to such models; third, operators must have infrastructure in place to disseminate the outputs of those models; and fourth, there must be a target audience that can be influenced by such content.

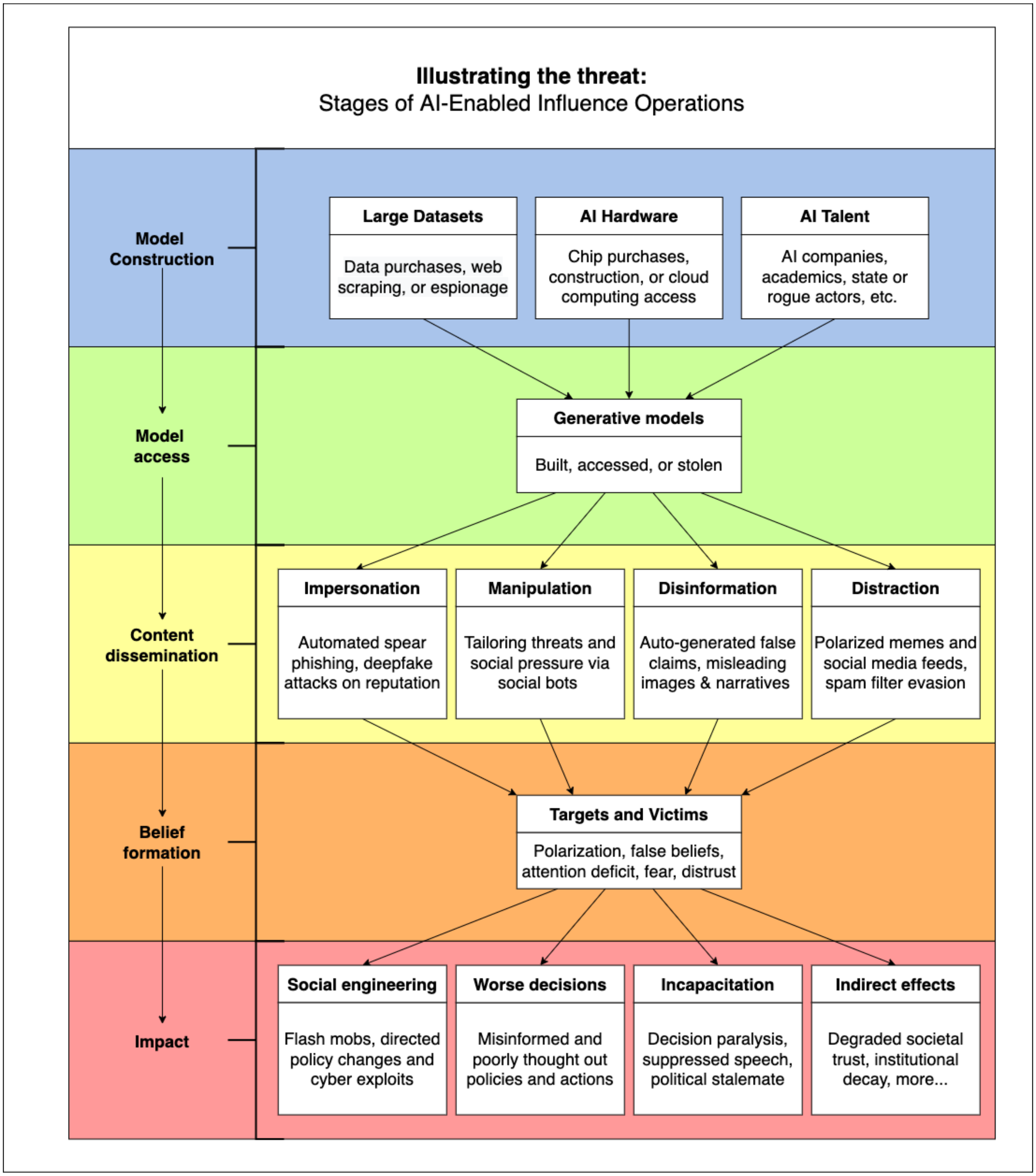
In Figure 4, we illustrate these points of intervention. For example, a threat actor can use generative model capabilities by accessing a model directly, building it themselves, or stealing the model. Any mitigation that intervenes at the **Model Access** stage should impact one or more of those three avenues.

For each of these stages, we can think about how an influence operation might be disrupted by using the following sets of questions as starting points:

- **Model Design and Construction:** How could AI models be built so they are robust against being misused to create disinformation? Could governments, civil society, or AI producers limit the proliferation of models capable of generating misinformation?
- **Model Access:** How could AI models become more difficult for bad actors to access for influence operations? What steps could AI providers and governments take?
- **Content Dissemination:** What steps can be taken to deter, monitor, or limit the spread of AI-generated content on social media platforms or news sites? How might the "rules of engagement" on the internet be altered to make the spread of AI-generated disinformation more difficult?

[132] For one example document that has compiled many strategies and resources for anti-misinformation campaigns, see Vivian Bianco et al., *Countering Online Misinformation Resource Pack* (UNICEF Regional Office for Europe and Central Asia, August 2020), https://www.unicef.org/eca/media/13636/file. See also Kalina Bontcheva et al., *Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression* (UNESCO, September 2020), https://en.unesco.org/publications/balanceact.

[133] There are other kill chain models that describe the ways disinformation operators conduct campaigns and how this process could be interrupted. See, for instance, Sedova et al., *AI and the Future of Disinformation Campaigns: Part 2: A Threat Model*; Bruce Schneier, "Toward an Information Operations Kill Chain," Lawfare, April 24, 2019, https://www.lawfareblog.com/toward-information-operations-kill-chain. However, for the purposes of analyzing the impact of AI language models specifically on disinformation, we use this simplified kill chain model.



Illustrating the threat:
Stages of AI-Enabled Influence Operations
Model Construction
Model access
Content dissemination
Belief formation
Impact
Large Datasets
Data purchases, web scraping, or espionage
AI Hardware
Chip purchases, construction, or cloud computing access
AI Talent
AI companies, academics, state or rogue actors, etc.
Generative models
Built, accessed, or stolen
Impersonation
Automated spear phishing, deepfake attacks on reputation
Manipulation
Tailoring threats and social pressure via social bots
Disinformation
Auto-generated false claims, misleading images & narratives
Distraction
Polarized memes and social media feeds, spam filter evasion
Targets and Victims
Polarization, false beliefs, attention deficit, fear, distrust
Social engineering
Flash mobs, directed policy changes and cyber exploits
Worse decisions
Misinformed and poorly thought out policies and actions
Incapacitation
Decision paralysis, suppressed speech, political stalemate
Indirect effects
Degraded societal trust, institutional decay, more...

Figure 4: Stages of intervention of AI-enabled influence operations. To disrupt a propagandist's use of language models for influence operations, mitigations can target four stages: (1) Model Design and Construction, (2) Model Access, (3) Content Dissemination, and (4) Belief Formation. Ultimately, intervening at these stages attempts to mitigate both the direct and indirect effects of influence operations.

- **Belief Formation:** If internet users are ultimately exposed to AI-generated content, what steps can be taken to limit the extent to which they are influenced?

We evaluate each mitigation by paying specific attention to four categories: (1) technical feasibility, (2) social feasibility, (3) downside risk, and (4) impact—four key considerations that stakeholders should use to assess the desirability of pursuing any particular mitigation. In more detail:

- **Technical feasibility** refers to the ability to implement a proposed mitigation on a technical level, without regard to social or political considerations. Some mitigations admit mature and low-cost technical solutions, while others require technical abilities that do not exist, are under question, or would require massive changes to existing technical infrastructure.
- **Social feasibility** refers to the political, legal, and institutional feasibility of a particular mitigation, assuming that it is technically possible to implement. The following questions serve as useful guides for assessing this metric: (1) Can the mitigation be successfully implemented unilaterally, without coordination across multiple independent actors? (2) Do the key actors who could implement the proposed mitigation have incentives in favor of doing so? (3) Would the proposed mitigation be actionable under existing law, regulation, and industry standards? Social feasibility will likely vary by region of interest.
- **Downside risk** refers to the negative impacts, including via negative externalities and second-order effects that a mitigation may cause. Notable downside risks that apply to multiple potential mitigations include heightened forms of censorship, the risk of the mitigation itself being politicized, and the risk of bias (such as inadvertently promoting certain perspectives, cultures, or languages over others).
- Finally, **impact** attempts to evaluate how effective a proposed mitigation would be at reducing the threat of AI-enabled influence operations. For instance, the mitigation "identify all AI-written text on the internet and remove it" is neither technically nor socially feasible, but if it could be implemented, this strategy would completely mitigate the effect of AI-powered influence operations (and thus have high impact). By contrast, "warn people about the dangers of AI-authored content" is much more feasible—but also far less impactful for reducing the effect of AI influence campaigns.

Of note, we do not attempt to separate this list of mitigations into "worth trying" and "fine to ignore" categories. Individual stakeholders capable of implementing any of these strategies must weigh the pros and cons of doing so. We also encourage additional research to address mitigations that fall outside of our model. We do not lay out mitigations that could shape the distribution of threat actor intentions (e.g., norm development, threats of retaliation) nor that could reduce harms that result from new beliefs shaped by a successful influence campaign. These warrant additional attention, but are not captured by our model.

In addition, we underscore that we discuss each mitigation in terms of **who or what institutions** would primarily be responsible for their implementation. But this leaves open the question of **why** these institutions would implement certain mitigations—specifically, whether they would do so voluntarily or should be compelled by regulators to take certain actions. By framing these mitigations in terms of the enacting institutions, we do not mean to suggest that this problem should be left to the voluntary actions

		Promise; if implemented...	Limitation
Model Design & Construction	AI Developers Build Models With More Detectable Outputs	Influence operations with language models will be easily discoverable	Technically challenging and requires coordination across developers
	AI Developers Build Models That Are More Fact-Sensitive	Language models will be less effective at spreading falsehoods	Technical methods are still being explored; may only impact some influence operations
	Developers Spread Radioactive Data to Make Generative Models Detectable	Makes it easier to detect if content is AI generated	Technically uncertain and may be easily circumvented
	Governments Impose Restrictions on Training Data Collection	Limits creation of new models (but only for those in jurisdictions that comply)	Data access restrictions would require high political will
	Governments Impose Access Controls on AI Hardware	Prevents some future models from being developed altogether	Restrictions on semiconductors could escalate geopolitical tensions and hurt legitimate businesses
Model Access	AI Providers Impose Stricter Usage Restrictions on Models	Makes it more difficult for propagandists to obtain cutting-edge models for campaigns	Requires coordination across AI providers and risks hurting legitimate applications
	AI Providers Develop New Norms Around Model Release	Restricts access to future models, but unlikely to prevent propagandists from obtaining already-public ones	Requires coordinating across AI providers and could concentrate capabilities among a small number of companies
	AI Providers Close Security Vulnerabilities	Prevents misuse and access of models via theft and tampering	Only affects one route to model access
Content Dissemination	Platforms and AI Providers Coordinate to Identify AI Content	Increases the likelihood of detecting AI-enabled influence operations	Will not affect platforms that do not engage; may not work in encrypted channels
	Platforms Require "Proof of Personhood" to Post	Increases the costs of waging influence operations	Current proof of personhood tests are often gameable by determined operators
	Entities That Rely on Public Input Take Steps to Reduce Their Exposure to Misleading AI Content	Protects entities relying on public inputs from AI-enabled campaigns	Significant changes to public comment systems could disincentivize participation
	Digital Provenance Standards Are Widely Adopted	Increases detection of AI-generated content	Significant changes would require large-scale coordination
Belief Formation	Institutions Engage In Media Literacy Campaigns	Mitigates the impact of influence operations	May reduce trust in legitimate content
	Developers Provide Consumer-Focused AI Tools	Increases the likelihood of people consuming high quality information	AI tools may be susceptible to bias; users may become overly reliant on them

Table 9: Summary of Example Mitigations and Selected Promise/Limitation

of AI developers and social media platforms. Updated regulations may be called for, and future research could unpack whether government intervention is needed (or desirable) for various mitigations. While we expect mitigations to be applicable across different countries, we focus below specifically on the United States to substantiate our points.

5.2 Model Design and Construction

The first stage at which key stakeholders could attempt to disrupt the spread of AI-powered disinformation is when language models are initially conceptualized and trained. How could these models be built differently (or how could they be limited from being built at all) such that it would become harder down the line to use them in influence operations? While the following mitigations might be useful, it is important to emphasize that the ability to construct these models is rapidly proliferating, as discussed in Section 4. Since most of these mitigations only affect the development of individual models—and getting consensus on any of these mitigations across all AI developers with the capability of constructing large language models will be very difficult—they generally score low on the metric of social feasibility.

The most reliable method for ensuring that large language models are not used in influence operations is to simply not build large language models. Every other proposed change to the design and construction of these models will be less effective at preventing misuse than not building the model in the first place. However, a complete stop to the development of new large language models is extremely unlikely, and so we focus primarily in this section on how these models could be built *differently* to reduce the risk of misuse.

5.2.1 AI Developers Build Models With More Detectable Outputs

Detecting AI-generated outputs of language models is currently a hard problem that is only getting harder as models improve.[134] However, some actions might be taken based on experiences in other AI subfields to increase the detectability of model outputs. In the subfield of computer vision, researchers at Meta have demonstrated that images produced by AI models can be identified as AI-generated if they are trained on "radioactive data"—that is, images that have been imperceptibly altered to slightly distort the training process. This detection is possible even when as little as 1% of a model's training data is radioactive and even when the visual outputs of the model look virtually identical to normal images.[135] It may be possible to build language models that produce more detectable outputs by similarly training them on radioactive data; however, this possibility has not been extensively explored, and the approach may ultimately not work.[136]

Rather than training on radioactive data, statistical perturbations might be introduced to a model's output by directly manipulating its parameters, thereby distinguishing its outputs from normal text and

[134] This is especially true for human detection. For example, researchers found a consistent trend that larger models produce text that is harder to distinguish from human written text; see Brown et al., "Language Models are Few-Shot Learners."

[135] Alexandre Sablayrolles et al., "Radioactive data: tracing through training," *37th International Conference on Machine Learning, ICML 2020* PartF168147-11 (February 3, 2020): 8296–8305, https://doi.org/10.48550/arxiv.2002.00937.

[136] There has been some success demonstrating that radioactive data can be used to induce certain types of behavior in language models; see Eric Wallace et al., "Concealed Data Poisoning Attacks on NLP Models," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021, 139–150, https://doi.org/10.48550/arxiv.2010.12563. However, it is not clear whether radioactive data can be used to generate models whose outputs can be reliably attributed to them.

making detection easier. Past research has identified tools that can be used to detect statistical patterns in outputs from less advanced models such as GPT-2; however, as models become bigger and develop a richer understanding of human text, these detection methods break down if the parameters of the models themselves are not deliberately perturbed in order to enable detection.[137]

However, there are reasons to think that it is difficult to build either highly detectable language models or reliable detection models. Linguistic data—especially across relatively short snippets of text—is already more compressed than in images, with far less room to express the subtle statistical patterns that the Facebook researchers relied on to detect AI-generated images. Still, it is possible that research could identify methods to statistically "fingerprint" a language model.[138] But it is unlikely that individual social media posts will ever be attributable directly to an AI model unless such fingerprints are sufficiently sophisticated: if the patterns permitting such detection were possible, they risk being clear enough for operators to screen out.[139] However, these strategies for building more detectable models may still make it possible to attribute larger-scale corpora of text to specific models, though this remains an open question.

Even if some models are designed or redesigned to produce outputs that are traceable at sufficient sizes, attackers could simply gravitate toward other models that are not similarly manipulated. For this mitigation to have a significant impact, it would require high levels of coordination across AI developers who have the ability to deploy large language models. Adversaries with the capability to create their own large language models may merely face additional costs, rather than a loss of capability. Furthermore, operating models that detect whether text is AI-generated represents a challenge, as these will have to be frequently updated to be reliable.

[137] On detection, see Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, July 2019, 111–116, https://doi.org/10.18653/V1/P19-3019. However, similar statistical methods perform less well for larger models such as GPT-3 and GROVER; see Leon Fröhling and Arkaitz Zubiaga, "Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover," *PeerJ Computer Science* 7 (April 6, 2021): 1–23, ISSN: 23765992, https://doi.org/10.7717/peerj-cs.443. In addition, none of this research assumes a realistic, adversarial threat model, in which attackers are aware that their posts are being assessed to potentially attribute machine authorship. Under this more realistic scenario, attackers could deploy very easy countermeasures, such as altering temperature settings to sample from a wider distribution of possible outputs in order to evade detection.

[138] Tao Xiang et al., "Protecting Your NLG Models with Semantic and Robust Watermarks," *arxiv:2112.05428 [cs.MM]*, December 10, 2021, https://doi.org/10.48550/arxiv.2112.05428.

[139] As an example of a trivially circumventable strategy, AI developers could embed special "zero-width" characters in the outputs of their models, which would not immediately be visible to users but which would easily be spotted by automated monitoring tools. There is some research into the use of zero-width characters to attack large language models—see Nicholas Boucher et al., "Bad Characters: Imperceptible NLP Attacks," *2022 IEEE Symposium on Security and Privacy*, June 2022, 1987–2004, ISSN: 10816011, https://doi.org/10.48550/arxiv.2106.09898; Luca Pajola and Mauro Conti, "Fall of Giants: How popular text-based MLaaS fall against a simple evasion attack," *Proceedings - 2021 IEEE European Symposium on Security and Privacy, Euro S and P 2021*, April 2021, 198–211, https://doi.org/10.48550/arxiv.2104.05996-but little research into their use as a defensive strategy, in large part because an attacker who was aware that such characters were being inserted into model outputs could easily just remove them before posting content online.

Criteria	Assessment
Technical Feasibility	It is an open technical question whether developers will be able to build models that produce detectable outputs.
Social Feasibility	To be implemented effectively, detectable models would require input and coordination across deployers of large language models, which may be socially infeasible.
Downside Risk	There are few obvious downside risks to developing detectable models, assuming there is a low false-positive rate.
Impact	If most or all models are detectable, then influence operations with language models will be easily discoverable.

5.2.2 AI Developers Build Models That Are More Fact-Sensitive

The dominant paradigm in natural language generation emphasizes "realism" in text generation over other possible values. Models are trained to generate text that effectively mimics (some subsample of) human text, without inherent regard for the truthfulness of the claims that it makes.[140] This means that false claims that are commonly believed may be just as likely for a model to produce as true claims under the current dominant approach to training language models.[141]

It may be possible to train AI models in such a way that they are incentivized to make more factually grounded claims, which could produce models that carry less risk of producing falsehoods even if they were accessible to bad actors.[142] Significant progress has been made in this area by training models that make use of web searches to improve the factual content of their responses, or that use reinforcement learning techniques to reward more factually correct responses—though these approaches embed their own set of biases about which claims count as "true" or "correct."[143] Other methods attempt to modify the text output to be well-supported by evidence.[144] While these methods are far from perfect, they can significantly reduce the risk that language models will produce misinformation during ordinary usage.

Nonetheless, most successful influence operations include, or build from, claims that have a kernel of truth.[145] Even a language model that produced no false claims could still be used to produce politically slanted or unfalsifiable statements, to shift public attention and discourse, or to engineer false beliefs due to selective context and inauthentic authorship. In fact, in the hands of the right operator, a model that stuck closely to the truth in its outputs might be more persuasive than a model that frequently lied.

[140] For instance, language models trained on large quantities of internet text will be trained on a large amount of fiction, which can lead them to substitute creative writing for facts.

[141] True claims are often a narrow target. Large language models such as GPT-3 are not necessarily truthful by default. See Buchanan et al., *Truth, Lies, and Automation: How Language Models Could Change Disinformation*.

[142] Owain Evans et al., "Truthful AI: Developing and governing AI that does not lie," *arxiv:2110.06674*, October 13, 2021, https://doi.org/10.48550/arxiv.2110.06674.

[143] Evans et al.; Ryan Lowe and Jan Leike, "Aligning Language Models to Follow Instructions," OpenAI Blog, January 27, 2022, https://openai.com/blog/instruction-following/; Jacob Hilton et al., "WebGPT: Improving the Factual Accuracy of Language Models through Web Browsing," OpenAI Blog, December 16, 2021, https://openai.com/blog/webgpt/.

[144] Hannah Rashkin et al., "Measuring Attribution in Natural Language Generation Models," *arxiv:2112.12870 [cs.CL]*, August 2, 2022, https://doi.org/10.48550/arxiv.2112.12870.

[145] As Starbird, Arif, and Wilson write, "To be effective, a disinformation campaign must be based around a 'rational core' of plausible, verifiable information or common understanding that can be reshaped with disinformation—for example half-truths, exaggerations, or lies." See Kate Starbird, Ahmer Arif, and Tom Wilson, "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," *Proceedings of the ACM on Human-Computer Interaction Vol: CSCW, Article 127*, CSCW 2019, ISSN: 25730142, https://doi.org/10.1145/3359229.

And further, if this mitigation did meaningfully make it harder for propagandists to misuse language models, it would still require coordination across AI developers to ensure that malicious actors do not simply gravitate toward models that were not trained using similar methods. Finally, to be up to date with the current state of the world, models might have to be retrained very frequently—a requirement that may impose prohibitive costs.

Criteria	**Assessment**
Technical Feasibility	AI developers are exploring ways to make models more fact sensitive, with promising signs of improvement.
Social Feasibility	For the mitigation to be fully implemented, it would require a high degree of coordination between developers of models.
Downside Risk	If language models are more truthful, they may be more persuasive and in turn inadvertently improve the persuasive capabilities of propagandists.
Impact	More truthful language models may be less likely to spread blatant misinformation, but can still serve influence operations relying on true, non-falsifiable, or politically slanted content.

5.2.3 Developers Spread Radioactive Data to Make Generative Models Detectable

Above, we described that AI developers could attempt to insert "radioactive data" into their datasets when training language models in order to create more detectable outputs. A drawback of this approach is that it requires significant coordination—radioactive data must be inserted by each developer into their own training pipeline. Alternatively, AI researchers, media companies, or governments themselves could choose to proliferate radioactive data directly onto the internet, in locations where it would likely be scooped up by any organization hoping to train a new language model.[146] This would require far less coordination and could potentially make AI outputs more detectable for all future language models. However, this would not affect models that have already been trained, and may be ineffective if developers take steps to filter their training data—a procedure that is common when training models.

This strategy would require proliferators to engage in secretive posting of large amounts of content online, which raises strong ethical concerns regarding the authority of any government or company to deliberately reshape the internet so drastically. In addition, this mitigation would only affect language models trained in the same language in which the radioactive data itself was written. It is also unclear how much of the internet would need to be "radioactive" in this way to meaningfully affect models. And, perhaps most importantly, it remains deeply unclear if this approach would actually result in models with more detectable outputs, for the reasons discussed previously in Section 5.2.1. It seems likely that, even with the use of radioactive training data, detecting synthetic text will remain far more difficult than detecting synthetic image or video content.

[146] Similar proposals have been advanced in the domain of visual deepfakes, as a way of increasing the likelihood that synthetic images produced from the most common models will be detectable to defenders. Hwang, *Deepfakes: A Grounded Threat Assessment*.

Criteria	Assessment
Technical Feasibility	While approaches to inserting radioactive data exist for images, it is unclear if this would work for text.
Social Feasibility	A well-resourced actor could unilaterally spread radioactive content that would likely be included in training data for future models.
Downside Risk	Large-scale, secret proliferation of data online raises significant concerns about the desirability of any one group changing the distribution of content on the internet so drastically.
Impact	It is unclear whether this retraining would result in more detectable outputs, and thus detectable influence operations.

5.2.4 Governments Impose Restrictions on Data Collection

The basis of any large language model is a vast quantity of training data in the form of text generated by real humans. While some of this data is typically taken from relatively structured sources such as Wikipedia, a large majority of data usually comes from tools like Common Crawl that scrape the web for publicly available text.[147] Regulatory or legal changes that would make this type of scraping more difficult to conduct might slow the growth of large language models, while simultaneously forcing developers to focus on extracting information from more structured sources.[148]

These changes could be grounded in changes to federal data privacy laws. Regulations that require internet users to be informed about what their personal data is used for—such as the General Data Protection Regulation (GDPR) in the EU—may slow down large language model development.[149] At the extreme end, governments could try to prohibit organizations from mass scraping the web for content at all. More targeted measures could aim at improving cybersecurity for personalized data on social media

[147] CommonCrawl freely publishes its archives of web data. See "So you're ready to get started.," Common Crawl, accessed June 27, 2022, https://commoncrawl.org/the-data/get-started/. But anyone can build their own software for web scraping or use other tools to extract data from websites.

[148] This would in turn have two follow-on effects: learning language from more factually grounded, more formal sources like online news or encyclopedia articles might make models more likely to produce true statements, while also making them significantly less capable of mimicking the language of highly specific target demographics. On using data restrictions to make language models more truthful, see Evans et al., "Truthful AI: Developing and governing AI that does not lie": 63.

[149] Article 14 of the GDPR requires companies that engage in web scraping of personal information regarding EU citizens to inform data subjects that their personal information has been collected and to grant them certain rights regarding the use of their data. See Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. L 119/1, art. 14. Major exemptions to this requirement do exist that would likely protect the scraping of textual data for the purposes of scientific research into language models (see ibid., art. 14(5)(b)); however, it is less clear to what extent GDPR may force companies looking to develop commercial AI models to identify impacted data subjects and expressly inform them of their inclusion in a training dataset. Due to the possibility of membership inference attacks on models that could be used to infer personal information about EU citizens, other components of the GDPR relating to protection of personal data may also be implicated in situations where AI developers use web scraping to create training datasets. For research into membership inference, see Nicolas Papernot et al., "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, October 2016, https://doi.org/10.48550/arxiv.1610.05755; and Reza Shokri et al., "Membership Inference Attacks against Machine Learning Models," *Proceedings - IEEE Symposium on Security and Privacy*, October 2016, 3–18, ISSN: 10816011, https://doi.org/10.48550/arxiv.1610.05820. At minimum, at least one company has been fined for non-compliance with Article 14 of the GDPR; see "Poland: First GDPR fine triggers controversial discussions," ePrivacy Blog, May 17, 2019, https://blog.eprivacy.eu/?p=544. This suggests that even if GDPR does not actually prohibit data scraping (including of personal information) for the purposes of language model construction, companies may feel that it is necessary to spend significantly more on lawyers and compliance efforts to avoid running afoul of the law.

platforms or prohibiting foreign acquisition of major platforms.[150]

These mitigations are significantly out of step with the current regulatory environment in the United States, which has not yet passed any comprehensive data privacy laws.[151] The Supreme Court has also recently ruled that scraping publicly available data from the web, even in violation of a terms of service agreement, does not violate the Computer Fraud and Abuse Act, the primary cybersecurity law in the United States.[152] Moreover, comprehensive data privacy laws that significantly affect the ability of language model developers to collect data may have large effects in other industries, while also having an uncertain ability to constrain developers outside of the United States. If implemented poorly, data protection measures may harm researchers' ability to detect and develop countermeasures against influence campaigns more than they hinder campaign planners.[153]

Beyond language models, it may be more feasible to regulate the collection or resale of image or video data. Specific state-level laws, like the Illinois Biometric Information Privacy Act (BIPA), restrict the ability of AI developers to scrape specific types of data—most often pictures of private individuals' faces—without informed consent.[154] Such laws have occasionally resulted in successful legal action against AI developers, as when the ACLU successfully used BIPA to compel Clearview AI to screen out data from Illinois residents in its model training pipeline and to sharply limit access to its facial recognition tools within Illinois.[155] Limiting access to relevant training data can meaningfully disrupt the creation of models that can later be used maliciously; at the same time, to the extent that such limitations are possible at all, they will likely be feasible only for certain restricted sets of training data, such as social media posts or images of private individuals' faces.

Criteria	Assessment
Technical Feasibility	Governmental policy to penalize data collection is likely possible without technical innovation; however, preventing access to internet-based training data is likely difficult.
Social Feasibility	More extreme forms of data access restrictions would require high political will.
Downside Risk	Limiting training data will negatively harm legitimate industries that may rely on language models or their training data and could undermine future detection models.
Impact	Without restricting data collection for all actors, impact is likely limited.

[150] See Todd C. Helmus and Marta Kepe, "A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda," *RAND Corporation*, June 2021, https://doi.org/10.7249/RR-A894-1; Chapter 1 in Eric Schmidt et al., *Final Report* (National Security Commission on Artificial Intelligence, 2021), https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf#page=52, 50, 405; and Austin Mooney, "Spotlight On Sensitive Personal Data As Foreign Investment Rules Take Force," *National Law Review* 11, no. 163 (February 18, 2020), https://www.natlawreview.com/article/spotlight-sensitive-personal-data-foreign-investment-rules-take-force.

[151] Thorin Klosowski, "The State of Consumer Data Privacy Laws in the US (And Why It Matters)," *New York Times*, September 6, 2021, https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/.

[152] Supreme Court of the United States, "Van Buren v. United States," October 2020, https://www.supremecourt.gov/opinions/20pdf/19-783_k53l.pdf.

[153] Nadya Bliss et al., "An Agenda for Disinformation Research," *arxiv:2012.08572 [cs.CY]*, December 2020, https://doi.org/10.48550/arxiv.2012.08572.

[154] Biometric Information Privacy Act, 740 Ill. Comp. Stat. § 14/1–25 (2008).

[155] ACLU v. Clearview AI, Inc., 2020 CH 04353 (Cir. Ct. Cook City., Ill.).

5.2.5 Governments Impose Controls on AI Hardware

Another path toward limiting the construction of large language models involves either limiting access to or monitoring the usage of AI hardware.[156] This could be achieved in a number of ways, including restrictions on the amount of computing power that individual organizations can use to train AI models, disclosure requirements for all AI projects requiring more than a certain threshold of computing power, or export controls on specialized chips.

Monitoring computing power usage may be difficult; some estimates suggest that a model 200 times larger than the current largest language model could be trained using less than 0.5% of worldwide cloud computing resources.[157] Even if major expenditures of computing power could reliably be identified and tracked, this power is a highly general resource; there is currently little way to tell that an organization purchasing a large amount of computing power is planning to train a large language model as opposed to, say, running climate simulations. However, increasing differentiation between AI compute and non-AI compute could make this easier in the future.[158]

Monitoring for large models is currently a difficult task, but semiconductor manufacturing equipment (SME) export controls or restrictions on access to cloud computing resources are easier to implement. In October 2022, the US government announced export controls on semiconductors, SMEs, and chip design software directed at China.[159] These controls could slow the growth in computing power in China, which may meaningfully affect their ability to produce future language models. Extending such controls to other jurisdictions seems feasible as the semiconductor supply chain is extremely concentrated.[160] Another (not mutually exclusive) restriction could involve mandating (or cloud computing companies could voluntarily implement) approval processes for projects requiring enough computing power to build a sophisticated language model. Even simply mandating stock and flow accounting of high-end AI chips could help identify which actors are capable of producing large language models.

To be effective, export controls on computing hardware need to be properly enforced and handle cases such as stockpiling of chips, re-exports via other jurisdictions, and so on. Computing hardware restrictions could also incentivize nation-states to accelerate their indigenous production of AI chips, though some reports argue that it is infeasible for China to scale up the domestic production of SME.[161] Furthermore, for the purpose of controlling language model development (or even AI development), export

[156] See, for example, Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," *arxiv:2004.07213 [cs.CY]*, April 2020, https://doi.org/10.48550/arxiv.2004.07213

[157] Andrew Lohn and Micah Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* (Center for Security and Emerging Technology, January 2022), https://doi.org/10.51593/2021CA009.

[158] As one example, AI training may use lower-precision chips; see Shar Narasimhan, "NVIDIA, Arm, and Intel Publish FP8 Specification for Standardization as an Interchange Format for AI," NVIDIA Technical Blog, September 14, 2022, https://developer.nvidia.com/blog/nvidia-arm-and-intel-publish-fp8-specification-for-standardization-as-an-interchange-format-for-ai/

[159] US Department of Commerce, Bureau of Industry, and Security, "Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC)," *Press Release*, October 7, 2022, https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file; US Department of Commerce, Bureau of Industry, and Security, "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification," *Docket No. 220930-0204, RIN 0694-AI94*, October 13, 2022, https://public-inspection.federalregister.gov/2022-21658.pdf.

[160] Saif M. Khan and Carrick Flynn, *Maintaining China's Dependence on Democracies for Advanced Computer Chips* (Center for Security and Emerging Technology, April 2020), https://cset.georgetown.edu/publication/maintaining-chinas-dependence-on-democracies-for-advanced-computer-chips/.

[161] Khan and Flynn.

controls on hardware are a blunt instrument and have far-reaching consequences on global trade and many non-AI industries.[162] Finally, it is worth keeping in mind that often the most impactful propagandists—governments themselves—are those with the capability to plausibly circumvent the hardware restrictions mentioned above.

Criteria	Assessment
Technical Feasibility	Some hardware-related controls would not require any technical innovation; however, this likely varies significantly.
Social Feasibility	Restrictions on semiconductors and SMEs have been applied to China; cloud computing restrictions could also be done unilaterally or voluntarily.
Downside Risk	Export controls on semiconductors or semiconductor manufacturing equipment could escalate geopolitical tensions and hurt legitimate businesses.
Impact	US export controls would largely affect the development of future language models in other jurisdictions.

5.3 Model Access

Once models are built, developers can choose how users interact with them. AI providers have some actions available to them that might reduce bad actors' access to generative language models. At the same time, these actions could be highly costly for organizations looking to commercialize their models and would require large amounts of cooperation across all relevant AI providers to ensure that propagandists could not simply gravitate toward other equally capable models without similar restrictions in place.

5.3.1 AI Providers Impose Stricter Controls on Language Models

As discussed in Section 2, the access regimes governing today's large language models generally fall into one of three categories: fully private, fully public, or private but accessible under restricted conditions, such as the use of gated API access. Access to many of the most powerful current large language models is partially available through APIs, which provides developers with a number of choices regarding potential access or use restrictions that could be imposed upon their models:

1. Developers could require potential users to submit the proposed purposes for which they intend to use a model, and revoke access if actual usage appears to diverge too far from this proposal. This type of restriction was originally a core component of OpenAI's API access regime, though it has since been replaced with a faster, more automated sign-up process.[163]

2. Even if the above proposal is adopted, individuals granted API access may often seek to build applications—for instance, chatbots—that give other end users the ability to indirectly input text

[162] Jordan Schneider and Irene Zhang, "New Chip Export Controls and the Sullivan Tech Doctrine with Kevin Wolf," ChinaTalk, October 11, 2022, https://www.chinatalk.media/p/new-chip-export-controls-explained.

[163] Bryan Walsh, "OpenAI's GPT-3 gets a little bit more open," Axios, November 18, 2021, https://www.axios.com/2021/11/18/openai-gpt-3-waiting-list-api.

to a model. These types of applications may indirectly expose the model to bad actors. Developers could therefore impose access restrictions that forbid API users from creating applications that give other users the ability to input arbitrary text to the model.

3. Developers might choose to restrict model access to only trusted institutions, such as known companies and research organizations, and not to individuals or governments likely to use their access to spread disinformation. Huawei initially appears to have intended an access regime along these lines for its PanGu-α model.[164]

4. Developers could further limit the number of outputs that individual users can generate within a certain period of time, or they could require review of users who seem to be submitting anomalously large numbers of queries. This would limit the scale of influence operations that rely on language models, but might not prevent their use in more tailored cases (such as generating a smaller number of news articles).

5. Where API access is granted, developers might also impose restrictions on the types of inputs that users are allowed to submit. For instance, the image-generating model DALL•E 2 attempts to screen out user-submitted queries that are intended to produce "violent, adult, or political" outputs.[165] Such efforts may require significant effort to keep them up to date as new controversial issues arise.

This does not represent an exhaustive list of potential access restrictions. All such restrictions, however, share certain downsides. First, effective restrictions may be difficult for developers to implement, especially if they require manual review or appeal processes. Second, organizations looking to commercialize their models have strong incentives to forego onerous review processes on potential customers. Third, user restrictions are only effective if enough institutions implement strong enough access restrictions to box out bad actors; otherwise, propagandists can simply gravitate toward models with less severe restrictions.

In other words, this proposed mitigation has the makings of a classic collective action problem: the most effective outcome requires coordination across multiple actors, each of whom has incentives to default. In addition, the proposal can only be effective so long as there are no publicly released models that are as effective and easy to use as those maintained by AI developers behind API restrictions. However, if public models are sufficient for propagandists, then this mitigation will likely be less effective.

Despite these limitations, strong industry norms—including norms enforced by industry standards or government regulation—could still make widespread adoption of strong access restrictions possible. As long as there is a significant gap between the most capable open-source model and the most capable API-controlled model, the imposition of monitoring controls can deny hostile actors some financial benefit.[166] Cohere, OpenAI, and AI21 have already collaborated to begin articulating norms around access to large language models, but it remains too early to tell how widely adopted, durable, and forceful these guidelines will prove to be.[167]

[164] Wiggers, "Huawei trained the Chinese-language equivalent of GPT-3."

[165] "Curbing Misuse at Dall-E 2," OpenAI, accessed June 27, 2022, https://openai.com/dall-e-2/.

[166] For a quantitative justification as to why, even if there are good public models available, restrictions on access to (better) private models can still impose non-negligible costs on propagandists, see Musser, "A Cost Analysis of Generative Language Models and Influence Operations."

[167] "Best Practices for Deploying Language Models," Cohere, June 2, 2022, https://txt.cohere.ai/best-practices-for-deploying-language-models/.

Finally, there may be alternatives to APIs as a method for AI developers to provide restricted access. For example, some work has proposed imposing controls on who can use models by only allowing them to work on specialized hardware—a method that may help with both access control and attribution.[168] Another strand of work is around the design of licenses for model use.[169] Further exploration of how to provide restricted access is likely valuable.

Criteria	Assessment
Technical Feasibility	Some AI developers already restrict usage of models behind APIs.
Social Feasibility	Limiting how AI providers' language models are used reflects a collective action problem: it requires coordination across AI providers, each of whom has an incentive to defect.
Downside Risk	Limiting access concentrates more power in the hands of a few AI providers and risks undermining those who could benefit from model use.
Impact	If AI developers are governed by norms of restricted use, it could mitigate the potential of AI-enabled influence operations. However, this assumes comparable open-source model developers do not exist.

5.3.2 AI Providers Develop New Norms Around Model Release

Traditionally, AI researchers have felt bound by what Thomas Merton referred to as the "communism of the scientific ethos," a norm that holds that a willingness to share information in the interests of full and open collaboration is integral to the scientific enterprise.[170] This norm is not merely a behavioral quirk of scientists; the free and open flow of information is critical for the advancement of science and technology as a whole, and progress in AI has long rested on strong norms of openness and collaboration. But as AI models become increasingly lucrative, this norm is challenged by a competing instinct to privatize models and data in order to commercialize them. In addition, norms of openness in AI research are challenged by safety concerns associated with powerful models that open up new attacks, including the scalable epistemic attacks made possible by powerful language models.[171]

Norms regarding data sharing and model release are currently in flux, largely due to progress in large language models. OpenAI has twice broken previous norms regarding model release, first by choosing to delay a full release of GPT-2 in order "to give people time to assess the properties of these models, discuss their societal implications, and evaluate the impacts of release after each stage," and then again a year later by choosing not to release GPT-3 at all, instead commercializing it behind an API paywall.[172] Both of these decisions drew serious criticism at the time, though the use of an API in lieu of a full model release now appears to be somewhat common among AI providers capable of producing cutting-edge

[168] Huili Chen et al., "DeepAttest: An end-to-end attestation framework for deep neural networks," *Proceedings of the 46th International Symposium on Computer Architecture*, June 2019, 487–498, ISSN: 10636897, https://doi.org/10.1145/3307650.3322251.

[169] "Responsible AI Licenses (RAIL)," Responsible AI Licenses (RAIL), accessed September 14, 2022, https://www.licenses.ai/.

[170] Robert K. Merton and Norman W. Storer, *The Sociology of Science: Theoretical and Empirical Investigations* (Univ. of Chicago Press, 1973).

[171] Percy Liang et al., "The Time Is Now to Develop Community Norms for the Release of Foundation Models," 2022, https://crfm.stanford.edu/2022/05/17/community-norms.html.

[172] Alex Radford et al., "Better Language Models and Their Implications," OpenAI Blog, February 14, 2019, https://openai.com/blog/better-language-models/.

language models.[173] In the domain of text-to-image models, a sitting member of Congress recently urged the US National Security Advisor and the acting director of the Office of Science and Technology Policy to address the "unsafe releases" of text-to-image models that do not have content restrictions, because they have been used to generate dangerous images.[174]

While we do not make specific claims about the substance of desirable research norms, a growing body of research is dedicated to examining the types of norms that could be developed to govern AI research, especially in the sphere of large language models. These norms could include staged release of models, the adoption of tradeoff frameworks to assess the risks of open-sourcing models, mechanisms for accepting public feedback and reports of misuse, and prepublication safety review.[175] Implementing any of these norms may require new institutional mechanisms, such as a Partnership on AI-style[176] organization for natural language processing researchers, the creation of a clear set of principles around issues like data collection and model release for large language models, or formal principles regarding what type of risk assessment is expected of AI developers prior to model release.[177] These institutional mechanisms could help solidify new norms around model design, model release, and model access and would have the potential to significantly impact the ability of propagandists to make use of large language models.

Criteria	Assessment
Technical Feasibility	This mitigation does not require technical innovation.
Social Feasibility	The development of norms around language model release for cutting-edge models requires coordination, and open-source developers may choose to ignore those norms.
Downside Risk	Norms that restrict model release may concentrate know-how in the hands of a smaller number of AI providers and impede beneficial AI progress.
Impact	The mitigation would be useful for restricting access to current and future cutting-edge models, but this is unlikely to prevent propagandists from gaining access to already-public models.

5.3.3 AI Providers Close Security Vulnerabilities

Actors seeking to make use of AI-generated content for propaganda may not be constrained by formal access restrictions to relevant models and research. They may employ covert espionage to steal models and information that will enable construction of their own models, or they may aim to engage in

[173] Jeremy Howard, "Some thoughts on zero-day threats in AI, and OpenAI's GPT-2," fast.ai, February 15, 2019, https://www.fast.ai/posts/2019-02-15-openai-gp2.html; "OpenAI Trains Language Model, Mass Hysteria Ensues," Approximately Correct, February 17, 2019, https://www.approximatelycorrect.com/2019/02/17/openai-trains-language-model-mass-hysteria-ensues/; "Microsoft's First GPT-3 Product Hints at the Commercial Future of OpenAI," TNW, June 5, 2011, https://thenextweb.com/news/microsofts-first-gpt-3-product-hints-commercial-future-openai-syndication.

[174] "Representative Anna Eshoo to Jake Sullivan and Alondra Nelson," September 20, 2020, https://eshoo.house.gov/sites/eshoo.house.gov/files/9.20.22LettertoNSCandOSTPonStabilityAI.pdf.

[175] Irene Solaiman et al., "Release Strategies and the Social Impacts of Language Models," *arxiv:1908.09203 [cs.CL]*, August 2019, https://doi.org/10.48550/arxiv.1908.09203; Aviv Ovadya and Jess Whittlestone, "Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning," *arxiv:1907.11274 [cs.CY]*, July 2019, https://doi.org/10.48550/arxiv.1907.11274.

[176] "Partnership on AI," Partnership on AI, accessed October 29, 2022, https://partnershiponai.org/.

[177] For one example of risk assessment for synthetic media, see "C2PA Specifications: C2PA Harms Modelling," Coalition for Content Provenance and Authenticity, accessed September 14, 2022, https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html.

cyberattacks or other forms of sabotage that allow them to manipulate the outputs of already existing language models.[178] For instance, language model poisoning or supply chain attacks on AI providers may allow adversaries to output propaganda from language models they do not possess—manipulating them from afar.[179] Similarly, threat actors may also seek to obtain access to cutting-edge, non-public generative models through human vulnerabilities and insider threats at AI institutions.

By developing or hiring groups to simulate adversary attempts to gain access to cutting-edge model capabilities, AI providers can identify and reduce vulnerabilities. Such red-teaming exercises should search not just for cybersecurity vulnerabilities, but also ways in which insider threats or mathematically sophisticated attacks on the AI training process could result in compromised models. Such red teaming can inform a holistic assessment on the risk of the model being misused or applied to produce propaganda. However, while red teaming may successfully identify some vulnerabilities, it is unlikely that all can be caught, and for many types of vulnerabilities that appear to be inherent in modern AI systems, it is unclear how successful any form of technical mitigation can be. Moreover, closing security vulnerabilities is only useful in the context of AI models that have not been made publicly available, as propagandists looking to make use of public models would not need to surreptitiously steal or compromise such models.

Criteria	Assessment
Technical Feasibility	Some red-teaming exercises can be performed today, but some defense methods for protecting valuable cyber assets remain research problems.
Social Feasibility	Individual AI developers can implement this mitigation unilaterally.
Downside Risk	There are no obvious downside risks.
Impact	Closing security vulnerabilities is useful insofar as future models are superior for propaganda purposes than already-public models.

5.4 Content Dissemination

AI-generated content is ultimately only a threat if it reaches and influences real human beings. In general, the interventions most likely to slow the spread of AI-generated propaganda may be those that could be successful against all propaganda, AI-generated or not. However, in this section, we briefly outline a few mitigations that might specifically manage to slow the spread of AI-authored content.

5.4.1 Platforms and AI Providers Coordinate to Identify AI Content

It is not clear how companies should respond if or when they judge that content on their platforms was generated by a language model. There are a wide number of plausibly legitimate use cases for AI-generated content on social media, including brand chatbots designed to provide customer service,

[178] For a taxonomy of the progression of machine learning vulnerabilities to adversarial influence and a series of case studies on these threats, see "MITRE | ATLAS," MITRE, accessed October 29, 2022, https://atlas.mitre.org/.

[179] For instance, in a "model spinning" attack, a threat actor can modify the model to output manipulated narratives whenever a user inputs an adversary-selected trigger word, all without compromising performance. See Bagdasaryan and Shmatikov, "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures." For a general overview of the types of attacks that can be used to target the mathematical peculiarities of AI systems, see Andrew Lohn, *Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity* (Center for Security and Emerging Technology, December 2020), https://doi.org/10.51593/2020CA006.

comedy bots meant to mimic or parody specific authors, or auto-generated news announcements.[180] For this reason, it is unlikely that social media platforms would choose to simply issue a blanket ban on AI-generated content.[181]

Even if platforms do not issue blanket bans, they might still build in rules regarding appropriate uses of language models into their terms of service. Should accounts generating automated content be required to publicly disclose the origin of content they post? Should posts determined to have been authored by an AI be flagged?[182] If platforms know that certain external sites host AI-generated content—especially content of a political nature—without disclosing it as such, might that be in itself sufficient grounds to block links to those sites? All of these interventions could be plausible ways to reduce the spread of AI-generated misinformation—assuming it can be identified as such.

Actually detecting content that comes from an AI model, however, is not trivial. Without the aid of AI developers, social media platforms trying to identify machine authorship would be restricted to merely looking for statistical patterns in text and user metadata.[183] Current tools for this do not provide the level of confidence that would likely be required for platforms to take disruptive action against accounts, do not work on texts the length of a typical social media post, and are likely to perform worse as models improve.[184]

However, collaboration between platforms and AI companies may make detection of larger-scale campaigns using AI generation more feasible. For instance, model owners might store outputs so that they

[180] Some of these types of uses already exist; for instance, the account dril_gpt2 on Twitter (https://twitter.com/dril_gpt2) uses GPT-2 to generate tweets in the style of the dadaist Twitter comedian dril.

[181] Some social media companies have restrictive policies around the posting of AI-generated images, but even these policies are usually only applicable in certain cases—most commonly, when there is an (assumed) intent to deceive behind the production of the image. See, for instance, Monica Bickert, "Enforcing Against Manipulated Media," *Meta*, January 6, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/, which contains the following explicit exemption: "This policy does not extend to content that is parody or satire, or video that has been edited solely to omit or change the order of words." The same type of reasons that have led social media companies to avoid adopting blanket bans on AI-generated visual content will also make blanket bans on AI-generated text content unlikely.

[182] The impact of flagging content as AI-generated on audiences' belief formation processes is unknown and may be unintuitive; in one study, for instance, researchers found that survey respondents were just as likely to view "AI-generated" profiles of Airbnb hosts as trustworthy, compared to human-authored profiles. However, when respondents were told that some profiles were human-authored and some were AI-generated, they viewed the profiles they believed were AI-generated as less trustworthy than human-authored profiles. Maurice Jakesch et al., "AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness," *CHI '19: Proceedings of CHI Conference on Human Factors in Computing Systems*, May 2019, https://doi.org/10.1145/3290605.3300469.

[183] Humans and machine learning-based detection systems differ in their respective competencies, and can currently perform better at detection together by covering each other's blindspots. See Daphne Ippolito et al., "Automatic Detection of Generated Text is Easiest when Humans are Fooled," *arXiv:1911.00650 [cs.CL]*, July 2020, 1808–1822, https://doi.org/10.48550/arXiv.1911.00650.

[184] One possible statistical method for identifying AI-generated text is provided by Hendrik Strobelt and Sebastian Gehrmann, "Catching a Unicorn with GLTR: A tool to detect automatically generated text," *gltr.io*, accessed October 29, 2022, http://gltr.io/. But this method assumes that language models will sample text from a relatively constrained distribution, such that the likelihood of unpredictable word patterns ends up significantly lower than is observed in authentic human text. As language models become larger, however, they become capable of accurately modeling a larger distribution of text, decreasing the risk that they will fall into noticeable "most-likely-next-word" ruts. Additionally, many language models permit users to directly manipulate a "temperature" setting, which directly serves to sample from a more unpredictable range of next word outputs when generating text, thereby evading this detection tool more directly.

can be traced back to the users who generated them.[185] Social media companies could then flag content on their platforms that they suspect may be inauthentic and work with AI companies to determine if any was generated by a language model. This type of collaboration could have follow-on benefits: once an AI company ascertains that a user is reposting outputs to social media, they can work with platforms to determine if other content generated by that user has been reposted to other social media platforms, potentially catching other coordinated inauthentic accounts that the platforms may initially have missed.

This strategy would miss content that is posted to encrypted social media platforms, such as WhatsApp channels. In addition, disinformation is also posted to social media platforms that do not support robust search features and are unlikely to cooperate with AI companies to monitor content, such as Parler and Gab, though it may still be possible to scan public posts on these sites for potential AI-generated content.[186] Without collaboration from the platforms themselves, this mitigation strategy may have only a limited impact.

Despite these drawbacks, partnerships between platforms and AI companies have certain advantages. Unlike imposing onerous up-front access restrictions, this type of monitoring is less likely to alienate potential users from signing up for API access to a language model, which may make it more attractive to AI companies. While bad actors may want to avoid using AI services that engage in this type of monitoring, AI companies can more easily maintain some secrecy about how they monitor for reposted content, making it harder to evade monitoring mechanisms.

Criteria	Assessment
Technical Feasibility	Versions of this implementation may be feasible for monitoring publicly posted content, but may be infeasible for encrypted social media channels.
Social Feasibility	Coordination between AI developers and social media companies requires a significant number of bilateral partnerships.
Downside Risk	There are few obvious downside risks assuming the detection models are accurate. If not, they risk flagging the wrong accounts.
Impact	The impact depends on the extent of collaboration between platforms and AI companies, and will not cover all social platforms.

5.4.2 Platforms Require “Proof of Personhood” to Post

Current policies regarding social media usage range from not requiring any form of registration to requiring that accounts be affiliated with real names and unique email addresses, and, at times, requiring

[185] This strategy will necessarily be imperfect, as propagandists can always make small or trivial changes to model outputs before posting them to social media. If detection relies on hash matches, operators may easily evade detection by doing so. However, not all operators may be savvy enough to realize that detection makes use of hashes, so this strategy may still have some usefulness. Relying on close-but-not-exact matches to output text, by contrast, introduces a higher level of statistical uncertainty in attribution, though at sufficient scales, campaigns with slightly altered text could still be linked to the use of an AI model with meaningful confidence.

[186] For analyses of Parler and Gab, including an overview of the extent of their content moderation practices, see David Thiel et al., *Contours and Controversies of Parler* (Stanford Internet Observatory, 2021), https://fsi.stanford.edu/publication/contours-and-controversies-parler and David Thiel and Miles McCain, *Gabufacturing Dissent: An in-depth analysis of Gab* (Stanford Internet Observatory, 2022), https://cyber.fsi.stanford.edu/publication/gabufacturing-dissent-an-in-depth-analysis-of-gab.

users to submit “video selfies” for proof of personhood.[187] However, any of these approaches can be circumvented by malicious actors: they can register many “burner” email addresses to create fake accounts and hire inexpensive labor to complete proof of humanness checks.

Platforms could, however, more uniformly require higher standards of proof of personhood in order to verify that content is not being produced by an AI and reposted to their sites. This could involve requiring more reliable forms of authentication when users sign up for an account, for instance, by asking a user to take a live video of themselves posing, or asking for some alternative form of biometrics. Alternatively, platforms could require users to occasionally pass tests to demonstrate humanness before posting content; these tests could either be administered randomly, at periodic intervals, or when a particular user is posting at a high volume. CAPTCHAs are one way to demonstrate humanness in this way; however, a determined adversary can cheaply circumvent them. Outside of tests, another proposed approach includes decentralized attestation of humanness.[188]

This mitigation would not make it impossible for propagandists to copy-paste content from a language model into a social media platform and post it. Instead, it would be meant to disrupt operational setups that rely on bots that directly query and post content from language models without explicit human intervention. While this may only describe a minority of influence operations, having such a fully automated capability might be useful to propagandists; for instance, an account could be configured to query a language model every few hours or days for an anodyne post with the intention of posting it directly to a social media platform. Operators would then need only log in every so often to post more explicitly political content, having fully automated the problem of enmeshing those political posts in a more realistic-seeming environment of unrelated content. Requiring checks to post content could meaningfully disrupt this type of operational setup.

There are several significant limitations to this mitigation, including potential infringements on privacy, limits to the types of operations it would mitigate, and limits to its effectiveness against operations by determined adversaries. First, from a privacy perspective, user authentication requirements would likely face resistance from users who are accustomed to an expectation of anonymity online, including users who hold such expectations for very legitimate reasons. Second, hummanness verifications are designed to address operations that rely on social media accounts to spread generated content, but do not affect other information channels—like email or fake news websites. Third, as mentioned above, for well-resourced actors like the Internet Research Agency, the costs of proof of humanness requirements may not be meaningful deterrents: purchasing a new SIM card or hiring cheap outsourced labor to pass a video test will not prevent these campaigns.

Finally, this mitigation introduces an underexplored potential for backlash: If platforms include a proof of humanness check, and propagandists pass such a check, the successful completion could increase the perceived credibility of the account—increasing the persuasive effect from the account in question. Future research could address this question directly.

[187] “Why you might be asked to upload a video selfie to confirm your identity on Instagram,” Facebook Help Centre, accessed October 29, 2022, https://m.facebook.com/help/1053588012132894.

[188] “The Internet Of Humans,” Proof Of Humanity, accessed October 29, 2022, https://www.proofofhumanity.id/.

Criteria	Assessment
Technical Feasibility	Various forms of human authentication have been piloted (and implemented) already.
Social Feasibility	Social media platforms and other websites can implement this mitigation unilaterally.
Downside Risk	More extreme forms of this mitigation would undermine online anonymity, which can stifle speech and undermine other human rights.
Impact	The impact depends on the specific implementation: basic CAPTCHA-like tests are gameable, but more novel implementations may increase costs of waging AI-enabled influence campaigns.

5.4.3 Entities That Rely on Public Input Take Steps to Reduce Their Exposure to Misleading AI Content

Many entities in society rely on public input for feedback, evidence of group beliefs, and legitimacy. For example, when making decisions that affect the community, local planning commissions often seek public comment to make informed decisions.[189] Similarly, private firms often ask for feedback on products, and media outlets often ask for tips on the issues of the day. The processes that these entities use for public comment constitute potential vectors for the abuse of language models to generate "comments" from the public in order to sway policymakers, local officials, or private entities.

Indeed, there have already been cases in which mass inauthentic comment campaigns have been identified in the US government, most notably when various technology companies submitted millions of comments to the FCC in 2017 regarding net neutrality, falsely using real customers' names to provide a veneer of legitimacy to the comments.[190] Comments generated by a large language model would be more difficult to identify as coordinated, since the comments in the FCC case followed a standard output and merely swapped synonyms for one another. As such, some level of reform to mechanisms for soliciting public input may be called for.

At the lowest end, this reform could simply involve making entities that solicit public comment more aware of the potential for inauthentic content being submitted that poses as public opinion. At the same time, this may have negative externalities: priming policymakers to be suspicious of public input, for example, may itself undermine democratic responsiveness.[191] Organizations soliciting public input might instead choose to implement stronger methods than common CAPTCHAs to ensure that public comments are authentic; currently, many US agencies simply assume that comments are legitimate and

[189] In the US context, each branch of the US government has mechanisms for soliciting input from members of the public. For Congress, the most common form of input is constituent calls or emails to their representatives; for the judicial system, the amicus brief provides a means for non-parties to a case to comment on its merits; and for executive agencies, the period of public comment required by the Administrative Procedures Act (APA) allows agencies to understand how affected parties might view proposed regulations.

[190] Jon Brodkin, "ISPs Funded 8.5 Million Fake Comments Opposing Net Neutrality," Wired, May 8, 2021, https://www.wired.com/story/isps-funded-85-million-fake-comments-opposing-net-neutrality/.

[191] Steve Balla et al., *Mass, Computer-Generated, and Fraudulent Comments* (Report to the Administrative Conference of the U.S., June 17, 2020), https://regulatorystudies.columbian.gwu.edu/mass-computer-generated-and-fraudulent-comments-0.

perform no follow-up on submitted comments.[192] Here, entities inviting comment will have to ensure that attempts to prevent AI-generated comments do not create frictions that prevent members of the public from participating.[193]

Criteria	Assessment
Technical Feasibility	Basic defenses—like user authentication—to prevent bots from overwhelming public comment boards already exist.
Social Feasibility	Policy change will likely require coordination across multiple parts of government.
Downside Risk	Significant changes may disincentivize members of the public from participating in public comment periods.
Impact	The impact varies depending on the specific implementation, but could make public input solicitation much more robust.

5.4.4 Digital Provenance Standards Are Widely Adopted

Because technical detection of AI-generated text is challenging, an alternate approach is to build trust by exposing consumers to information about how a particular piece of content is created or changed. Tools such as phone cameras or word processing software could build the means for content creators to track and disclose this information.[194] In turn, social media platforms, browsers, and internet protocols could publicize these indicators of authenticity when a user interacts with content.

This intervention requires a substantial change to a whole ecosystem of applications and infrastructure in order to ensure that content retains indicators of authenticity as it travels across the internet. To this end, the Coalition for Content Provenance and Authenticity (C2PA) has brought together software application vendors, hardware manufacturers, provenance providers, content publishers, and social media platforms to propose a technical standard for content provenance that can be implemented across the internet.[195] This standard would provide information about content to consumers, including its date of creation, authorship, hardware, and details regarding edits, all of which would be validated with cryptographic signatures.[196]

Theoretically, this standard would work for AI-generated content, particularly if AI-as-a-service compa-

[192] Committee on Homeland Security U.S. Senate Permanent Subcommittee on Investigations and Governmental Affairs, *Abuses of the Federal Notice-and-Comment Rulemaking Process* (2019), https://tinyurl.com/5bamt57s; "Federal Rulemaking: Selected Agencies Should Fully Describe Public Comment Data and Their Limitations," *U.S. GAO*, September 2021, https://www.gao.gov/products/gao-21-103181. The GAO study found that, for some agencies, as many as 30% of individuals whose email addresses were associated with public comments reported not having written the comment submitted under their name. Many other agencies did not require email addresses or other types of identifying information for submitted comments, significantly reducing the ability of the agency to authenticate the identity of the commenter.

[193] In the US context, a stronger version could be that the APA itself is amended to mandate some level of vetting for the authenticity of public comments, or criminal liability could be imposed for institutions found to be impersonating members of the public. We do note, however, that the Administrative Conference of the United States (ACUS) has so far preferred not to propose any sweeping changes to the period for public comment. In part, this is because ACUS believes that AI-generated comments could have valuable use cases in the public comment process, such as by generating summaries of public comments or lowering barriers to submitting public comments. See Balla et al., *Mass, Computer-Generated, and Fraudulent Comments*

[194] For one example of a media provenance pipeline from certified authoring tools to browser extensions for verification, see Paul England et al., "AMP: Authentication of Media via Provenance," *MMSys 2021 - Proceedings of the 2021 Multimedia Systems Conference*, June 2021, 108–121, https://doi.org/10.48550/arxiv.2001.07886.

[195] "C2PA Specifications: C2PA Harms Modelling."

[196] "Verifiable Credentials Data Model v1.1," W3C, March 3, 2022, https://www.w3.org/TR/vc-data-model/.

nies opt in to self-declare authorship for each piece of content and require applications or individuals accessing their services through API to do the same. Over time, users may learn to trust the content that has provenance markers and distrust content that lacks them. However, these protocols cannot authenticate preexisting legacy content. In addition, while these measures can provide greater transparency about the creation, history, and distribution of files—including images and text files generated by word processing applications—they cannot provide a means for authenticating and tracking the spread of *raw text*, which can be copied and pasted from file to file without leaving a record in a specific file's history. To authenticate text provenance widely would require radical changes to internet protocols. For example, it is possible that the HTTP protocol would have to be modified to embed content provenance information. Since language models output raw text and not files, simply storing provenance information in files is sharply limited in its ability to help track the spread of AI-generated misinformation. More low-level changes may be needed to maximize the impact of this intervention.

If the provenance information for a piece of content contains information about the user, then this intervention would raise privacy risks.[197] This implementation could threaten anonymous speech on the internet. However, if only information to distinguish AI and human-generated content is added, then the privacy risks are lower.

Criteria	Assessment
Technical Feasibility	Promising technical paths exist, but the technology has not yet been proven.
Social Feasibility	Some progress has been made in coordinating between interested parties, but robust versions of this mitigation would require massive coordination challenges.
Downside Risk	Adding author information raises privacy risks.
Impact	Radical changes to guarantee content provenance would have high impact, but more feasible options would likely have limited impact.

5.5 Belief Formation

The preceding mitigations address the supply of AI-generated misinformation. However, as long as target audiences remain susceptible to propaganda that aligns with their beliefs, there will remain an incentive for influence operations generally, as well as incentives more specifically for propagandists to leverage AI to make those operations more effective. In this section, we therefore discuss two interventions that might help address the demand side of the misinformation problem: media literacy campaigns, and the use of AI tools to aid media consumers in interpreting and making informed choices about the information they receive.

[197] For more discussion of privacy risks here, see "Ticks or it didn't happen," WITNESS Media Lab, December 2019, https://lab.witness.org/ticks-or-it-didnt-happen/.

5.5.1 Institutions Engage in Media Literacy Campaigns

There is some evidence that media literacy campaigns can increase individuals' ability to discern between real and fake news online.[198] Existing media literacy tools that teach people how to "spot" coordinated accounts online, however, sometimes emphasize traits or mistakes that AI tools can avoid making, such as repetitiveness or a lack of "personal" content interspersed with more political content.[199] If current programs become outdated, media literacy will require updating. For example, if language models overcome repetition and lack of "personal" content, literacy campaigns can still combat the goals of the propagandists by teaching people to fact-check content in articles and to distinguish objective information from false, misleading, or slanted content.[200] These campaigns may have less impact, however, on distraction operations that crowd out genuine news.

Unlike many of the other mitigations listed above, the impact of media literacy campaigns is agnostic to human versus computer authorship. These efforts focus on teaching people how to analyze content, not necessarily to spot AI-generated content. Another form of digital literacy campaigns could be to teach people about AI-generated content specifically. If new "telltale" signs can be identified that represent common indicators of AI-powered influence operations, then this mitigation could be beneficial. However, if the most that can be said of AI-powered operations is that they look more authentic than human-operated campaigns, then this strategy may be misplaced. Emphasizing that any account on the internet could be an AI-powered bot may make people more likely to simply dismiss arguments they disagree with as inauthentic and not worth paying attention to, thereby exacerbating societal division and polarization. Overemphasizing the prevalence and danger of misinformation online may ultimately serve the same goal that propagandists themselves are often trying to achieve: making people inherently distrustful of any information or argument that conflicts with their preexisting beliefs.[201]

Criteria	Assessment
Technical Feasibility	No technical innovation is required.
Social Feasibility	A variety of actors could unilaterally lead educational campaigns.
Downside Risk	Educating about the threat of AI-enabled influence operations could reduce trust in genuine content or in online information environments more broadly.
Impact	Educational initiatives could help people distinguish reliable information from misinformation or slanted text, and mitigate the effects of influence operations (AI-generated or not).

[198] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren, "Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures," *Harvard Kennedy School Misinformation Review* 1, no. 2 (February 2020), https://doi.org/10.37016//MR-2020-008; Andrew M. Guess et al., "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India," *PNAS* 117, no. 27 (July 2020): 15536–15545, ISSN: 10916490, https://doi.org/10.1073/pnas.1920498117; Se Hoon Jeong, Hyunyi Cho, and Yoori Hwang, "Media Literacy Interventions: A Meta-Analytic Review," *Journal of Communication* 62, no. 3 (June 2012): 454–472, ISSN: 0021-9916, https://doi.org/10.1111/J.1460-2466.2012.01643.X; Todd C. Helmus et al., "Russian Propaganda Hits Its Mark: Experimentally Testing the Impact of Russian Propaganda and Counter-Interventions," *RAND Corporation*, October 2020, https://doi.org/10.7249/RRA704-3

[199] For an existing example of a media literacy tool that teaches users the "telltale" signs of troll accounts, see "Spot The Troll," Clemson University Media Forensics Hub, https://spotthetroll.org/.

[200] For one example of the effectiveness of these measures, see Gordon Pennycook et al., "Shifting attention to accuracy can reduce misinformation online," *Nature* 592 (7855 2021): 590–595, ISSN: 1476-4687, https://doi.org/10.1038/s41586-021-03344-2.

[201] Karen Hao, "The biggest threat of deepfakes isn't the deepfakes themselves," *MIT Technology Review*, October 10, 2019, https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/.

5.5.2 Developers Provide Consumer-Focused AI Tools

Just as generative models can be used to generate propaganda, they may also be used to defend against it. Consumer-focused AI tools could help information consumers identify and critically evaluate content or curate accurate information. These tools may serve as an antidote to influence operations and could reduce the demand for disinformation. While detection methods (discussed in Section 5.2.1) aim to detect whether content is synthetic, consumer-focused tools instead try to equip consumers to make better decisions when evaluating the content they encounter.

Possibilities for such tools are numerous.[202] Developers could produce browser extensions and mobile applications that automatically attach warning labels to potential generated content and fake accounts, or that selectively employ ad-blockers to demonetize them. Websites and customizable notification systems could be built or improved with AI-augmented vetting, scoring, and ranking systems to organize, curate, and display user-relevant information while sifting out unverified or generated sources.[203] Tools and built-in search engines that merely help users quickly contextualize the content they consume could help their users evaluate claims, while lowering the risk of identifying true articles as misinformation.[204] Such "contextualization engines" may be especially helpful in enabling users to analyze a given source and then find both related high-quality sources and areas where relevant data is missing. By reducing the effort required to launch deeper investigations, such tools can help to align web traffic revenue more directly with user goals, as opposed to those of advertisers or influence operators.[205] Another proposal suggests using AI-generated content to educate and inoculate a population against misleading beliefs.[206]

Some of the most promising AI-enabled countermeasures may leverage state-of-the-art generative models themselves, to reshift the offense-defense balance in favor of information consumers.[207] As generative models get better at producing persuasive arguments that exploit viewer biases and blindspots, defensive generative models could be used to help users detect and explain flaws in tailored arguments or to find artifacts in manipulated images.[208] Generative models that help users find relevant information can also be trained how to "show their work" by citing sources that support their answers.[209] Such methods could serve as building blocks for future tools that augment a consumer's ability to critically evaluate

[202] For a variety of examples of consumer-focused tools that help users control the information they see, see *Combatting Online Harms Through Innovation, Report to Congress* (Federal Trade Commission, June 16, 2022), https://www.ftc.gov/reports/combatting-online-harms-through-innovation.

[203] A particularly successful example of a curation tool is Live Universal Awareness Map, which has done near real-time source aggregation on conflicts in Ukraine and Syria while aiming to filter out state-sponsored propaganda. On karma and reputation systems, see Seger et al., *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*; and Christian Johnson and William Marcellino, *Bad Actors in News Reporting: Tracking News Manipulation by State Actors* (RAND Corporation, November 2021), https://doi.org/10.7249/RRA112-21

[204] The issue of false positives—identifying quality sources as misleading or false—is common with social media fact-checking recommendation systems, which often superficially associate new accurate articles with prior false ones, or fail to differentiate between false claims and claims that are contingent, probabilistic, or predictive in nature.

[205] Aviv Ovadya, "'Contextualization Engines' can fight misinformation without censorship," *Medium*, May 26, 2022, https://aviv.medium.com/contextualization-engines-can-fight-misinformation-without-censorship-c5c47222a3b7.

[206] "Humor over Rumor: Combating Disinformation Around COVID-19 in Taiwan," Global Governance Futures, June 2020, accessed September 14, 2022, https://www.ggfutures.net/analysis/humor-over-rumor-combating-disinformation-around-covid-19-in-taiwan; Herriman et al., "Asked and Answered: Building a Chatbot to Address Covid-19-Related Concerns."

[207] By tailoring to serve the needs of individual information consumers, such tools could equip consumers with decision-informing capabilities that would otherwise be too risky to implement at the scale of an entire platform.

[208] Jan Leike et al., "AI-Written Critiques Help Humans Notice Flaws," *OpenAI Blog*, June 13, 2022, https://openai.com/blog/critiques/.

[209] Reiichiro Nakano et al., "WebGPT: Browser-assisted question-answering with human feedback," *arxiv:2112.09332 [cs.CL]*, June 1, 2022, https://doi.org/10.48550/arxiv.2112.09332.

information.

Consumer-focused tools may also go beyond the individual, with more expensive, AI-enabled intelligence services that offer tools to businesses, governments, and other organizations that aim to increase their awareness of, and improve their responses to, influence operations.

Despite their prospective benefits, AI tools will also present risks. They are likely to be susceptible to forms of social bias, just as current models are. Defensive generative models that are aligned with consumer incentives may also exacerbate confirmation bias, as consumers may prefer information that tailors to their preexisting biases. Social media companies may make it difficult or against their policies for externally developed tools to interface with their platforms, both to protect privacy and to sustain user engagement. While social media companies may be in a good position to provide their own defensive AI tools, the divergence between their interests and those of their users would likely exceed that of third-party tool providers. Accordingly, tools created by platforms could also serve to discourage more effective policy action and to justify disabling the use of third-party tools that aren't as aligned with platform objectives.[210]

More powerful versions of web-searching generative models may also pose new unique risks if their range of action and reinforcable behavior is not carefully constrained. For models that are capable of generating and inputting text queries within other websites to find more relevant results, the incentive to return useful results could reward fraudulent behavior (e.g., editing and returning Wikipedia results if there aren't good sources[211]). While many such specific imagined threats are highly unlikely, the potential impacts of defensive generative models on search engine traffic and the internet itself should be accounted for.

Overall, consumer-focused AI tools provide a variety of opportunities to head off the impact of influence operations that employ stronger generative models, but they will require high-quality implementation.

Criteria	Assessment
Technical Feasibility	Creating AI tools that help people reason or highlight factual inaccuracies is an ongoing research problem, but some promising directions exist.
Social Feasibility	Some progress could be achieved unilaterally by researchers or entrepreneurs, but coordination with social media platforms would be required for broader effect.
Downside Risk	AI tools may be susceptible to bias, and people could become overly reliant on them.
Impact	If implemented well, defensive AI tools could have a big impact in helping consumers form accurate beliefs.

[210] For example, the use of such tools could be used to impress Congress with a platform's efforts, and to make the argument that users already have plenty of options to seek out or control the information they are exposed to, even if in practice the tools are designed to discourage use.

[211] Nakano et al., "WebGPT: Browser-assisted question-answering with human feedback."

6 Conclusions

While each of the mitigations discussed above are important to weigh on their own merits, there are some crosscutting conclusions that we offer to policymakers trying to think through the problem of AI-powered influence operations. Our shared assessments of these mitigations lead to the following main conclusions:

1. Language models are likely to significantly impact the future of influence operations.
2. There are no silver bullets for minimizing the risk of AI-generated disinformation.
3. New institutions and coordination (like collaboration between AI providers and social media platforms) are needed to collectively respond to the threat of (AI-powered) influence operations.
4. Mitigations that address the supply of mis- or disinformation without addressing the demand for it are only partial solutions.
5. More research is needed to fully understand the threat of AI-powered influence operations as well as the feasibility of proposed mitigations.

6.1 Language Models Will Likely Change Influence Operations

As outlined in Section 4, language models have the potential to significantly affect how influence operations are waged in the future—including the actors waging these campaigns, the behaviors of the propagandists, and the content included.

> *Actors:* If generative models become widely accessible, it will drive down the cost of producing propaganda; in turn, those who have refrained from waging influence operations in the past may no longer be disinclined. Private PR and marketing firms may develop knowledge in how to most effectively integrate these models, and thus serve as a resource and scapegoat for political actors seeking to outsource their campaigns.
>
> *Behaviors:* Language models offer to change how influence operations are waged. They may be deployed for dynamic generation of responses, automated cross-platform testing, and other novel techniques. Although we described a few new possible behaviors in this report, we suspect propagandists will use these models in unforeseen ways in response to the defensive measures that evolve.
>
> *Content:* Language models will likely drive down the cost and increase the scale of propaganda generation. As language models continue to improve, they will be able to produce persuasive text—text that is difficult to distinguish from human-generated content—with greater reliability, reducing the need for skilled writers with deep cultural and linguistic knowledge of the target population.

Although we foresee these changes in the medium term, there is some speculation at play. The extent to which language models change the nature of influence operations is dependent on critical unknowns, including diffusion and accessibility, and various technical and social uncertainties. We do not yet know who will control these models, and how information environments—like social media platforms—will adapt in a world where models are widely available for use.

6.2 There Are No Silver Bullet Solutions

Section 5 discussed a large number of possible strategies for managing the threat of AI-generated influence operations. Unfortunately, no proposed mitigation manages to be simultaneously (1) technically feasible, (2) institutionally tractable, (3) robust against second-order risks, and (4) highly impactful. The fact that large language models are increasingly proliferating—both behind paid APIs and in the form of openly released models—currently makes it all but impossible to ensure that large language models are never used to generate disinformation.

This is not an excuse for defeatism. Even if responding to the threat is difficult, AI developers who have built large language models have a responsibility to take reasonable steps to minimize the harms of those models. By the same token, social media companies have a continuing obligation to take all appropriate steps to fight misinformation, while policymakers must seriously consider how they can help make a difference. But all parties should recognize that any mitigation strategies specifically designed to target AI-generated content will not fully address the endemic challenges.

Even if better policies can be adopted to govern the majority of language models, very few interventions will stop a well-resourced, non-cooperative state from constructing its own alternatives. One option for countries like the United States would be to soften immigration requirements for AI talent, which could concentrate the ability to produce language models in a few countries—though this too will be unlikely to fully stop a sufficiently motivated nation-state from developing high capability systems of their own.

6.3 Collective Responses Are Needed

Many of the mitigations discussed above might have a meaningful impact in reducing AI-generated influence campaigns, but only if new forms of collaboration are developed. Strong norms among the AI community—regarding either the release of models or the training methods used to develop them—could make it harder for the most common language models to be induced to generate disinformation. We have also suggested that if detection of AI-generated text will be feasible at all, it will likely require relatively large "batches" of outputted text in order to attribute. Collaboration between social media companies and AI companies may be necessary in order to curate and attribute large batches of potentially inauthentic content.

The current US response to influence operations is fractured: fractured among technology companies, fractured among academic researchers, fractured between multiple government agencies, and fractured on the level of collaboration between these groups. Social media companies have different approaches to whether (and how) to treat influence operations; academics lack relevant data to understand related issues; AI developers often lack sufficient expertise to understand potential abuses of the technologies they create, and responsibilities for influence operations are not clearly delineated to any single US department or agency. Policymakers should consider creating stronger mechanisms and incentives to ensure coordination across all relevant stakeholders.[212]

6.4 Mitigations Must Address Demand As Well As Supply

All else being equal, the fact that a particular post was authored by an AI does not in itself make the content of that post less truthful or more destabilizing than the same content would be coming from a human. While this paper has focused on mitigations that would disrupt the pipeline between large language models and influence operations, it is important to emphasize that many other mitigations can be implemented or further strengthened that aim to reduce the spread of false or biased information generally. Some social media platforms have already implemented a number of these mitigations—though often not equitably between English-speaking countries and other regions. But influence operations appear to be a new normal of online activity, and more effort to improve these mitigations is warranted.

It is equally important, however, to emphasize that mitigations that disrupt the supply of misleading information are ultimately only partial solutions if the demand for misleading information remains unchanged. While people rarely demand to be misinformed directly, information consumers often demand information that is cheap and useful for their goals—something influence operations can tailor to with greater freedom from the constraints of reality.

From a selfish perspective, ignorance is often rational: it is not possible to be informed on everything, gathering accurate information can be boring, and countering false beliefs may have social costs.[213] Similarly, consuming and sharing disinformation may be entertaining, attract attention, or help an individual gain status within a polarized social group. When the personal costs of effortful analysis exceed the personal benefits, the likely result will be lower-quality contribution to group decision-making (e.g., sharing disinformation, free riding, groupthink, etc.).

6.5 Further Research Is Necessary

Many of the properties of large generative models are not fully understood. Similarly, clarity is still missing regarding both the structure and the impacts of many influence operations, which are conducted in secret.

Clarity on the scale of the threat posed by influence operations continues to be elusive. Is the actual impact of such campaigns proportionate to the attention they receive in the popular imagination and press coverage? How effective are existing platform-based mitigations—such as friction measures designed to slow down the virality of content—at reducing the spread of misinformation? As it relates

[212] The National Security Commission on AI, the Aspen Institute, and a variety of others have recommendations for how to integrate government efforts to counter foreign-sourced influence campaigns. See Schmidt et al., *Final Report*; *The Weaponization of Information: The Need for Cognitive Security* (RAND Corporation, April 27, 2017); Fletcher Schoen and Christopher J. Lamb, *Deception, Disinformation, and Strategic Communications: How One Interagency Group Made a Major Difference* (Center for Strategic Research Institute for National Strategic Studies, June 2012), https://ndupress.ndu.edu/Portals/68/Documents/stratperspective/inss/Strategic-Perspectives-11.pdf; Matt Chessen, *The MADCOM future* (Atlantic Council, September 2017), https://www.atlanticcouncil.org/in-depth-research-reports/report/the-madcom-future/; Sedova et al., *AI and the Future of Disinformation Campaigns: Part 2: A Threat Model*.

[213] Anthony Downs, "An Economic Theory of Political Action in a Democracy," *Journal of Political Economy* 65, no. 2 (1957): 135–150, https://www.jstor.org/stable/1827369.

to influence operations with generative models specifically, future research should unpack the differential impact these technologies may have on different populations. For example, relevant factors include the languages various models output most persuasively, and the media and internet fluency in different communities. AI developers and researchers could reach out to communities likely to be impacted to better understand their risks and needs.

A number of technical issues are also currently ambiguous. The relationship between model size, length of fine-tuning, and overall performance or persuasiveness, for instance, is unclear. While it is generally true that larger, more heavily trained models perform better across a wide variety of tasks—including disinformation-related ones—it is not clear whether fine-tuning a smaller model can reliably make up that gap. How do these factors change between models primarily trained on large, well-represented languages like English and those with more capability to use less well-represented languages? On the mitigation side, the feasibility of detection methods remains ambiguous. Although it seems reasonable to assume that (1) attributing short pieces of content as AI-generated will remain impossible and (2) detection might become possible at much larger scales, it is hard to be more specific than this. What scales are necessary to enable detection? How much can perturbing models or training on radioactive data alter this necessary threshold? Furthermore, how realistic is it to train models in ways that reduce their likelihood of outputting misleading content to begin with?

Further research would also be useful to better understand, model, and clarify the decision-making of propagandists themselves. Detailed analyses of the relative gains that malicious actors can capture by incorporating generative models into their operations are also lacking. It is similarly unclear whether API restrictions on large language models meaningfully discourage operators from accessing certain services, and if they do, whether operators are able to simply gravitate toward open-source models without any loss of capability.[214]

Finally, this is a rapidly moving field where norms have not yet solidified. Should AI developers release or restrict their models? Should internet researchers publish observed tactics of propagandists or keep them secret? To what extent can platforms and AI developers form meaningful partnerships that can aid in the detection and removal of inauthentic content? At the broadest level, thoughtful engagement with all of these questions—both from people within the relevant industries and from neutral, third-party observers—is a critical necessity.

[214] Forthcoming work from some of the authors will attempt to partially address this. See Musser, "A Cost Analysis of Generative Language Models and Influence Operations"

References

Allyn, Bobby. “Deepfake video of Zelenskyy could be ‘tip of the iceberg’ in info war, experts warn.” NPR, March 16, 2022. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

AlQuraishi, Mohammed. “Machine learning in protein structure prediction.” *Current Opinion in Chemical Biology* 65 (December 2021): 1–8. ISSN: 1367-5931. https://doi.org/10.1016/J.CBPA.2021.04.005.

Altay, Sacha, Anne Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. “Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions.” *Journal of Experimental Psychology: Applied*, October 28, 2021. ISSN: 1939-2192. https://doi.org/10.1037/XAP0000400.

“API.” OpenAI. Accessed January 31, 2022. https://openai.com/api/.

August 2020 Coordinated Inauthentic Behavior Report. Meta, September 1, 2020. https://about.fb.com/news/2020/09/august-2020-cib-report/.

Ayyub, Rana. “I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me.” *Huffington Post*, November 21, 2018. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

Bagdasaryan, Eugene, and Vitaly Shmatikov. “Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures.” *2022 IEEE Symposium on Security and Privacy*, 2022, 769–786. https://doi.org/10.1109/SP46214.2022.9833572.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. “Assessing the Russian Internet Research Agency’s impact on the political attitudes and behaviors of American Twitter users in late 2017.” *PNAS* 117, no. 1 (January 7, 2020). https://doi.org/10.1073/pnas.1906420116.

Baker, Bowen, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. “Learning to Play Minecraft with Video PreTraining (VPT).” OpenAI Blog, June 23, 2022. https://openai.com/blog/vpt/.

Balla, Steve, Reeve Bull, Bridget Dooling, Emily Hammond, Michael Herz, Michael Livermore, and Beth Simone Noveck. *Mass, Computer-Generated, and Fraudulent Comments*. Report to the Administrative Conference of the U.S., June 17, 2020. https://regulatorystudies.columbian.gwu.edu/mass-computer-generated-and-fraudulent-comments-0.

Bateman, John, Elonnai Hickok, Laura Courchesne, Isra Thange, and Jacob N. Shapiro. *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.

“WudaoAI.” *Beijing Academy of Artificial Intelligence*. Accessed October 30, 2022. https://wudaoai.cn/model/.

“Best Practices for Deploying Language Models.” Cohere, June 2, 2022. https://txt.cohere.ai/best-practices-for-deploying-language-models/.

Bianco, Vivian, Sergiu Tomsa, Mario Mosquera Vasques, and Svetlana Stefanet. *Countering Online Misinformation Resource Pack*. UNICEF Regional Office for Europe and Central Asia, August 2020. https://www.unicef.org/eca/media/13636/file.

Bickert, Monica. "Enforcing Against Manipulated Media." *Meta*, January 6, 2020. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

Bliss, Nadya, Elizabeth Bradley, Joshua Garland, Filippo Menczer, Scott W. Ruston, Kate Starbird, and Chris Wiggins. "An Agenda for Disinformation Research." *arxiv:2012.08572 [cs.CY]*, December 2020. https://doi.org/10.48550/arxiv.2012.08572.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. "On the Opportunities and Risks of Foundation Models." *arxiv:2108.07258 [cs.LG]*, August 2021. https://doi.org/10.48550/arxiv.2108.07258.

Bond, Shannon. "AI-generated fake faces have become a hallmark of online influence operations." *NPR*, December 15, 2022. https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations.

Bontcheva, Kalina, Julie Posetti, Denis Teyssou Agence, France Presse, France Trisha Meyer, Sam Gregory, U S Clara Hanot, and Diana Maynard. *Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression*. UNESCO, September 2020. https://en.unesco.org/publications/balanceact.

Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Van Den Driessche, et al. "Improving language models by retrieving from trillions of tokens." *arxiv:2112.04426 [cs.CL]*, December 2021. https://doi.org/10.48550/arxiv.2112.04426.

Boucher, Nicholas, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. "Bad Characters: Imperceptible NLP Attacks." *2022 IEEE Symposium on Security and Privacy*, June 2022, 1987–2004. ISSN: 10816011. https://doi.org/10.48550/arxiv.2106.09898.

Brodkin, Jon. "ISPs Funded 8.5 Million Fake Comments Opposing Net Neutrality." Wired, May 8, 2021. https://www.wired.com/story/isps-funded-85-million-fake-comments-opposing-net-neutrality/.

Brooking, Emerson T., and Jacob Shapiro. "Americans Were Worried About the Wrong Threat." Atlantic, January 10, 2020. https://www.theatlantic.com/ideas/archive/2021/01/bigger-threat-was-always-domestic/617618/.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33 (May 2020). ISSN: 10495258. https://doi.org/10.48550/arxiv.2005.14165.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *arxiv:2004.07213 [cs.CY]*, April 2020. https://doi.org/10.48550/arxiv.2004.07213.

Buchanan, Ben, 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://doi.org/10.51593/2021CA003.

Buchanan, Ben, and Taylor Miller. *Machine Learning for Policy Makers: What It Is and Why It Matters*. Belfer Center for Science and International Affairs, June 2017. https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

"Building a TB Scale Multilingual Dataset for Language Modeling." Hugging Face BigScience. https://bigscience.huggingface.co/blog/building-a-tb-scale-multilingual-dataset-for-language-modeling.

"C2PA Specifications: C2PA Harms Modelling." Coalition for Content Provenance and Authenticity. Accessed September 14, 2022. https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html.

Chen, Huili, Cheng Fu, Bita Darvish Rouhani, Jishen Zhao, and Farinaz Koushanfar. "DeepAttest: An end-to-end attestation framework for deep neural networks." *Proceedings of the 46th International Symposium on Computer Architecture*, June 2019, 487–498. ISSN: 10636897. https://doi.org/10.1145/3307650.3322251.

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. "Evaluating Large Language Models Trained on Code." *arxiv:2107.03374 [cs.LG]*, July 14, 2021. https://doi.org/10.48550/arxiv.2107.03374.

Chesney, Robert, and Danielle Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (2019): 1753. https://doi.org/10.15779/Z38RV0D15J.

Chessen, Matt. *The MADCOM future*. Atlantic Council, September 2017. https://www.atlanticcouncil.org/in-depth-research-reports/report/the-madcom-future/.

"Chinese propagandists court South-East Asia's Chinese diaspora." Economist, November 20, 2021. https://www.economist.com/asia/2021/11/20/chinese-propagandists-court-south-east-asias-chinese-diaspora.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, et al. "Scaling Instruction-Finetuned Language Models." *arxiv:2210.11416 [cs.LG]*, October 20, 2022. https://doi.org/10.48550/arxiv.2210.11416.

Cohere. "About." Accessed January 31, 2022. https://docs.cohere.ai/api-reference/.

Combatting Online Harms Through Innovation, Report to Congress. Federal Trade Commission, June 16, 2022. https://www.ftc.gov/reports/combatting-online-harms-through-innovation.

Commerce, US Department of, Bureau of Industry, and Security. "Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC)." *Press Release*, October 7, 2022. https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file.

———. "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification." *Docket No. 220930-0204, RIN 0694-AI94*, October 13, 2022. https://public-inspection.federalregister.gov/2022-21658.pdf.

Council, National Intelligence. *Intelligence Community Assessment: Foreign Threats to the 2020 US Federal Elections*. National Intelligence Council, March 10, 2021. https://int.nyt.com/data/documenttools/2021-intelligence-community-election-interference-assessment/abd0346ebdd93e1e/full.pdf.

"Curbing Misuse at Dall-E 2." OpenAI. Accessed June 27, 2022. https://openai.com/dall-e-2/.

Delaney, Jack. “I’m a freelance writer. A Russian media operation targeted and used me.” *The Guardian*, September 4, 2020. https://www.theguardian.com/technology/2020/sep/04/russia-media-disinformation-fake-news-peacedata.

Dhingra, Bhuwan, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. “Time-Aware Language Models as Temporal Knowledge Bases.” *Transactions of the Association for Computational Linguistics* 10 (March 2022): 257–273. ISSN: 2307387X. https://doi.org/10.1162/tacl_a_00459.

Dill, Janina, Scott D. Sagan, and Benjamin A. Valentino. “Kettles of Hawks: Public Opinion on the Nuclear Taboo and Noncombatant Immunity in the United States, United Kingdom, France, and Israel.” *Security Studies* 31, no. 1 (2022): 1–31. ISSN: 15561852. https://doi.org/10.1080/09636412.2022.2038663.

Ding, Jeffrey, and Jenny Xiao. “Recent Trends in China’s Large-Scale Pre-Trained AI Models.” *(Working Paper)*.

DiResta, Renee, and Shelby Grossman. *Potemkin Pages & Personas: Assessing GRU Online Operations, 2014-2019*. Stanford Internet Observatory, 2019. https://cyber.fsi.stanford.edu/io/publication/potemkin-think-tanks.

DiResta, Renee, Shelby Grossman, Samantha Bradshaw, Karen Nershi, Khadeja Ramali, and Rajeev Sharma. “In Bed with Embeds: How a Network Tied to IRA Operations Created Fake “Man on the Street” Content Embedded in News Articles.” *Stanford Internet Observatory*, December 2, 2021. https://cyber.fsi.stanford.edu/io/publication/bed-embeds.

DiResta, Renee, Michael McFaul, and Alex Stamos. “Here’s How Russia Will Attack the 2020 Election. We’re Still Not Ready.” *The Washington Post*, November 15, 2019. https://www.washingtonpost.com/opinions/2019/11/15/heres-how-russia-will-attack-election-were-still-not-ready/.

DiResta, Renée, Shelby Grossman, and Alexandra Siegel. “In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies.” *Political Communication* 39, no. 2 (2021): 222–253. ISSN: 10917675. https://doi.org/10.1080/10584609.2021.1994065.

Downs, Anthony. “An Economic Theory of Political Action in a Democracy.” *Journal of Political Economy* 65, no. 2 (1957): 135–150. https://www.jstor.org/stable/1827369.

Earl, Jennifer, Thomas V. Maher, and Jennifer Pan. “The digital repression of social movements, protest, and activism: A synthetic review.” *Science Advances* 8 (October 2022): 8198. https://www.science.org/doi/pdf/10.1126/sciadv.abl8198.

Emelyanov, Anton, Tatiana Shavrina, Oleh Shliazhko, and Artem Snegirev. “Russian GPT-3 models.” GitHub. https://github.com/ai-forever/ru-gpts#readme.

England, Paul, Henrique S. Malvar, Eric Horvitz, Jack W. Stokes, Cédric Fournet, Rebecca Burke-Aguero, Amaury Chamayou, et al. “AMP: Authentication of Media via Provenance.” *MMSys 2021 - Proceedings of the 2021 Multimedia Systems Conference*, June 2021, 108–121. https://doi.org/10.48550/arxiv.2001.07886.

Evans, Owain, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. “Truthful AI: Developing and governing AI that does not lie.” *arxiv:2110.06674,* October 13, 2021. https://doi.org/10.48550/arxiv.2110.06674.

Farid, Hany. "Creating, Using, Misusing, and Detecting Deep Fakes." *Journal of Online Trust and Safety* 1, no. 4 (September 2022). ISSN: 2770-3142. https://doi.org/10.54501/JOTS.V1I4.56.

"Fine-tuning." OpenAI. Accessed June 2022. https://beta.openai.com/docs/guides/fine-tuning.

"Finetuning Generation Models." Cohere. Accessed June 2022. http://web.archive.org/web/20220621204451/https://docs.cohere.ai/finetuning-wiki/.

Finnemore, Martha, and Kathryn Sikkink. "International Norm Dynamics and Political Change." *International Organization* 52, no. 4 (1998): 887–917. https://www.jstor.org/stable/2601361.

Fisher, Max. "Disinformation for Hire, a Shadow Industry, Is Quietly Booming." *New York Times,* July 25, 2021. https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html.

François, Camille. *Actors, Behaviors, Content: A Disinformation ABC Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses*. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, September 2019. https://science.house.gov/download/francois-addendum.

Frenkel, Sheera. "Iranian Disinformation Effort Went Small to Stay Under Big Tech's Radar." *New York Times*, June 30, 2021. https://www.nytimes.com/2021/06/30/technology/disinformation-message-apps.html.

Fröhling, Leon, and Arkaitz Zubiaga. "Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover." *PeerJ Computer Science* 7 (April 6, 2021): 1–23. ISSN: 23765992. https://doi.org/10.7717/peerj-cs.443.

Ganguli, Deep, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, et al. "Predictability and Surprise in Large Generative Models." *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 2022, 1747–1764. https://doi.org/10.1145/3531146.3533229.

Gehrmann, Sebastian, Hendrik Strobelt, and Alexander M. Rush. "GLTR: Statistical Detection and Visualization of Generated Text." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, July 2019, 111–116. https://doi.org/10.18653/V1/P19-3019.

Geller, Tom. "Overcoming the Uncanny Valley." *IEEE Computer Graphics and Applications* 28, no. 4 (July-Aug. 2008): 11–17. ISSN: 02721716. https://doi.org/10.1109/MCG.2008.79.

Gleicher, Nathaniel, Margarita Franklin, David Agranovich, Ben Nimmo, Olga Belogolova, and Mike Torrey. *Threat Report: The State of Influence Operations 2017-2020*. Meta, May 2021. https://about.fb.com/news/2021/05/influence-operations-threat-report/.

Goldstein, Josh A. "Foreign Influence Operations in the Cyber Age." PhD diss., University of Oxford, 2021. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.840171.

Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. "Can AI write persuasive propaganda?" *(Working Paper)*.

Goldstein, Josh A., and Renée DiResta. "This salesperson does not exist: How tactics from political influence operations on social media are deployed for commercial lead generation." *Harvard Kennedy School Misinformation Review 3*, no. 5 (September 2022). https://doi.org/10.37016/MR-2020-104.

Goldstein, Josh A., and Renee DiResta. "China's Fake Twitter Accounts Are Tweeting Into the Void." *Foreign Policy*, December 15, 2021. https://foreignpolicy.com/2021/12/15/china-twitter-trolls-ccp-influence-operations-astroturfing/.

Goldstein, Josh A., and Shelby Grossman. "How disinformation evolved in 2020," January 4, 2021. https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. https://www.deeplearningbook.org/.

Graphika. *Posing as Patriots*. Graphika, June 2021. https://graphika.com/reports/posing-as-patriots.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363, no. 6425 (January 25, 2019): 374–378. ISSN: 10959203. https://doi.org/10.1126/science.aau2706.

Grossman, Shelby, Gil Baram, Josh A. Goldstein, and Carly Miller. *Staying Current: An Investigation Into a Suspended Facebook Network Supporting the Leader of the Palestinian Democratic Reform Current*. Stanford Internet Observatory, February 10, 2021. https://purl.stanford.edu/tk756wp5109.

Grossman, Shelby, Chris Giles, Cynthia N. M., Miles McCain, and Blair Read. "The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists." Stanford Internet Observatory, December 2, 2021. https://stacks.stanford.edu/file/druid:bt877dz8024/20211202-tz-twitter-takedown.pdf.

Grossman, Shelby, Khadija H., and Emily Ross. *Royal Sockpuppets and Handle Switching: How a Saudi Arabia-Linked Twitter Network Stoked Rumors of a Coup in Qatar*. Stanford Internet Observatory, October 2020. https://stacks.stanford.edu/file/druid:hp643wc2962/twitter-SA-202009.pdf.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *PNAS* 117, no. 27 (July 2020): 15536–15545. ISSN: 10916490. https://doi.org/10.1073/pnas.1920498117.

Hao, Karen. "The biggest threat of deepfakes isn't the deepfakes themselves." *MIT Technology Review*, October 10, 2019. https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/.

Heim, Lennart. "Estimating PaLM's training cost." .xyz Blog, April 5, 2022. https://blog.heim.xyz/palm-training-cost/.

Helmus, Todd C., and Marta Kepe. "A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda." *RAND Corporation*, June 2021. https://doi.org/10.7249/RR-A894-1.

Helmus, Todd C., James V. Marrone, Marek N. Posard, and Danielle Schlang. "Russian Propaganda Hits Its Mark: Experimentally Testing the Impact of Russian Propaganda and Counter-Interventions." *RAND Corporation*, October 2020. https://doi.org/10.7249/RRA704-3.

Hernandez, Danny, and Tom B. Brown. "Measuring the Algorithmic Efficiency of Neural Networks." *arxiv:2005.04305 [cs.LG]*, May 2020. https://doi.org/10.48550/arxiv.2005.04305.

Herriman, Maguire, Elana Meer, Roy Rosin, Vivian Lee, Vindell Washington, and Kevin G. Volpp. "Asked and Answered: Building a Chatbot to Address Covid-19-Related Concerns." *NEJM Catalyst Innovations in Care Delivery*, June 18, 2020. https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0230.

Hilton, Jacob, Suchi Balaji, Relichiro Nakano, and John Schulman. “WebGPT: Improving the Factual Accuracy of Language Models through Web Browsing.” OpenAI Blog, December 16, 2021. https://openai.com/blog/webgpt/.

Ho, Ed. “An Update on Safety.” Twitter Blogs, February 7, 2021. https://blog.twitter.com/en_us/topics/product/2017/an-update-on-safety.

Holtzman, Ari, Jan Buys, Leo Du, Maxwell Forbes, and Yejin Choi. “The Curious Case of Neural Text Degeneration.” *arxiv:1904.09751 [cs.CL]*, February 19, 2019. ISSN: 16130073. https://doi.org/10.48550/arxiv.1904.09751.

Howard, Jeremy. “Some thoughts on zero-day threats in AI, and OpenAI’s GPT-2.” fast.ai, February 15, 2019. https://www.fast.ai/posts/2019-02-15-openai-gp2.html.

“Humor over Rumor: Combating Disinformation Around COVID-19 in Taiwan.” Global Governance Futures, June 2020. Accessed September 14, 2022. https://www.ggfutures.net/analysis/humor-over-rumor-combating-disinformation-around-covid-19-in-taiwan.

Hwang, Tim. *Deepfakes: A Grounded Threat Assessment*. Center for Security and Emerging Technology, July 2020. https://doi.org/10.51593/20190030.

Ippolito, Daphne, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. “Automatic Detection of Generated Text is Easiest when Humans are Fooled.” *arXiv:1911.00650 [cs.CL]*, July 2020, 1808–1822. https://doi.org/10.48550/arXiv.1911.00650.

Jakesch, Maurice, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. “AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness.” *CHI ’19: Proceedings of CHI Conference on Human Factors in Computing Systems*, May 2019. https://doi.org/10.1145/3290605.3300469.

Jeong, Se Hoon, Hyunyi Cho, and Yoori Hwang. “Media Literacy Interventions: A Meta-Analytic Review.” *Journal of Communication* 62, no. 3 (June 2012): 454–472. ISSN: 0021-9916. https://doi.org/10.1111/J.1460-2466.2012.01643.X.

Johnson, Christian, and William Marcellino. *Bad Actors in News Reporting: Tracking News Manipulation by State Actors*. RAND Corporation, November 2021. https://doi.org/10.7249/RRA112-21.

Jowett, Garth, and Victoria O’Donnell. *Propaganda & Persuasion*. 6th ed. SAGE Publications, 2014. ISBN: 1483323528.

Kahembwe, Emmanuel, and Subramanian Ramamoorthy. “Lower Dimensional Kernels for Video Discriminators.” *Neural Networks* 132 (December 2020): 506–520. https://doi.org/10.1016/j.neunet.2020.09.016.

Kallberg, Jan, and Stephen Col. Hamilton. “US military must prepare for POW concerns in the deepfake era.” C4ISRNET, August 23, 2021. https://www.c4isrnet.com/opinion/2021/08/23/us-military-must-prepare-for-pow-concerns-in-the-deepfake-era/.

Keskar, Nitish Shirish, Bryan Mccann, Lav R Varshney, Caiming Xiong, Richard Socher, and Salesforce Research. “CTRL: A Conditional Transformer Language Model for Controllable Generation.” *arxiv:1909.05858 [cs.CL]*, September 2019. https://doi.org/10.48550/arxiv.1909.05858.

Khan, Saif M., and Carrick Flynn. *Maintaining China’s Dependence on Democracies for Advanced Computer Chips*. Center for Security and Emerging Technology, April 2020. https://cset.georgetown.edu/publication/maintaining-chinas-dependence-on-democracies-for-advanced-computer-chips/.

Khrushchev, Mikhail. “Yandex Publishes YaLM 100B. It’s the Largest GPT-Like Neural Network in Open Source.” Medium, June 23, 2022. https://medium.com/yandex/yandex-publishes-yalm-100b-its-the-largest-gpt-like-neural-network-in-open-source-d1df53d0e9a6.

King, Gary, Jennifer Pan, and Margaret E. Roberts. “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument.” *American Political Science Review* 111, no. 3 (2017): 484–501. https://doi.org/10.1017/S0003055417000144.

Klosowski, Thorin. “The State of Consumer Data Privacy Laws in the US (And Why It Matters).” *New York Times*, September 6, 2021. https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/.

Knight, Will. “AI-Powered Text From This Program Could Fool the Government.” Wired, January 15, 2021. https://www.wired.com/story/ai-powered-text-program-could-fool-government/.

Kreps, Sarah, and Doug Kriner. “The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment.” *(Working Paper)*.

Kreps, Sarah, R. Miles McCain, and Miles Brundage. “All the News That’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation.” *Journal of Experimental Political Science* 9, no. 1 (November 2022): 104–117. ISSN: 2052-2630. https://doi.org/10.1017/XPS.2020.37.

Kundu, Kishalaya. “Criminals Used AI To Clone Company Director’s Voice And Steal $35 Million.” Screen Rant, October 14, 2021. https://screenrant.com/ai-deepfake-cloned-voice-bank-scam-theft-millions/.

Kurenkov, Andrey. “Lessons from the GPT-4Chan Controversy.” The Gradient, June 12, 2022. https://thegradient.pub/gpt-4chan-lessons/.

Leike, Jan, Jeffrey Wu, Catherine Yeh, and William Saunders. “AI-Written Critiques Help Humans Notice Flaws.” *OpenAI Blog*, June 13, 2022. https://openai.com/blog/critiques/.

Liang, Percy, Rishi Bommasani, Kathleen A. Creel, and Rob Reich. “The Time Is Now to Develop Community Norms for the Release of Foundation Models,” 2022. https://crfm.stanford.edu/2022/05/17/community-norms.html.

Liang, Percy, Rob Reich, and et al. “Condemning the deployment of GPT-4chan.” Accessed July 22, 2022. https://docs.google.com/forms/d/e/1FAIpQLSdh3Pgh0sGrYtRihBu-GPN7FSQoODBLvF7dVAFLZk2iuMgoLw/viewform?fbzx=1650213417672418119.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.” *ACM Computing Surveys*, September 2021. https://doi.org/10.1145/3560815.

Lohn, Andrew. *Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity*. Center for Security and Emerging Technology, December 2020. https://doi.org/10.51593/2020CA006.

Lohn, Andrew, and Micah Musser. *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* Center for Security and Emerging Technology, January 2022. https://doi.org/10.51593/2021CA009.

Lohn, Andrew J., and Krystal A. Jackson. *Will AI Make Cyber Swords or Shields?* Center for Security and Emerging Technology, August 2022. https://doi.org/10.51593/2022CA002.

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. "TimeLMs: Diachronic Language Models from Twitter." *arxiv.2202.03829 [cs.CL]*, February 2022, 251–260. https://doi.org/10.48550/arxiv.2202.03829.

Lowe, Ryan, and Jan Leike. "Aligning Language Models to Follow Instructions." OpenAI Blog, January 27, 2022. https://openai.com/blog/instruction-following/.

M.A., Renee DiResta, Josh A. Goldstein, and Shelby Grossman. "Middle East Influence Operations: Observations Across Social Media Takedowns." *Project on Middle East Political Science*, August 2021. https://pomeps.org/middle-east-influence-operations-observations-across-social-media-takedowns.

Mandiant. *'Ghostwriter' Influence Campaign: Unknown Actors Leverage Website Compromises and Fabricated Content to Push Narratives Aligned with Russian Security Interests*. Mandiant. https://www.fireeye.com/content/dam/fireeye-www/blog/pdfs/Ghostwriter-Influence-Campaign.pdf.

Mansimov, Elman, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. "Generating Images from Captions with Attention." *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, November 9, 2015. https://doi.org/10.48550/arxiv.1511.02793.

"Marv the Sarcastic Chat Bot." OpenAI API. https://beta.openai.com/examples/default-marv-sarcastic-chat.

Mazarr, Michael J., Ryan Michael Bauer, Abigail Casey, Sarah Anita Heintz, and Luke J. Matthews. *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment*. RAND Corporation, October 2019. https://doi.org/10.7249/RR2714.

Merton, Robert K., and Norman W. Storer. *The Sociology of Science: Theoretical and Empirical Investigations*. Univ. of Chicago Press, 1973.

Metz, Rachel. "How a deepfake Tom Cruise on TikTok turned into a very real AI company." CNN, August 6, 2021. https://edition.cnn.com/2021/08/06/tech/tom-cruise-deepfake-tiktok-company.

"Microsoft's First GPT-3 Product Hints at the Commercial Future of OpenAI." TNW, June 5, 2011. https://thenextweb.com/news/microsofts-first-gpt-3-product-hints-commercial-future-openai-syndication.

"MITRE | ATLAS." MITRE. Accessed October 29, 2022. https://atlas.mitre.org/.

"ML-Enhanced Code Completion Improves Developer Productivity." Google AI Blog. Accessed July 28, 2022. https://ai.googleblog.com/2022/07/ml-enhanced-code-completion-improves.html.

Mooney, Austin. "Spotlight On Sensitive Personal Data As Foreign Investment Rules Take Force." *National Law Review* 11, no. 163 (February 18, 2020). https://www.natlawreview.com/article/spotlight-sensitive-personal-data-foreign-investment-rules-take-force.

"Moravec's paradox." Wikipedia. Accessed June 29, 2022. https://en.wikipedia.org/wiki/Moravec%5C%27s_paradox.

Mu, Zhaoxi, Xinyu Yang, and Yizhuo Dong. "Review of end-to-end speech synthesis technology based on deep learning." *arxiv:2104.09995 [cs.SD]*, April 2021. https://doi.org/10.48550/arxiv.2104.09995.

Murphy, Matt. "Someone trained an A.I. with 4chan. It could get worse." Slate, August 3, 2022. https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html.

"Muse API." PAGnol. https://muse.lighton.ai/home.

Musser, Micah. "A Cost Analysis of Generative Language Models and Influence Operations." *(Working Paper)*.

Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, et al. "WebGPT: Browser-assisted question-answering with human feedback." *arxiv:2112.09332 [cs.CL]*, June 1, 2022. https://doi.org/10.48550/arxiv.2112.09332.

Narang, Sharan, and Aakanksha Chowdhery. "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance." Google AI Blog, April 5, 2022. https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html.

Narasimhan, Shar. "NVIDIA, Arm, and Intel Publish FP8 Specification for Standardization as an Interchange Format for AI." NVIDIA Technical Blog, September 14, 2022. https://developer.nvidia.com/blog/nvidia-arm-and-intel-publish-fp8-specification-for-standardization-as-an-interchange-format-for-ai/.

"NAVER Unveils HyperCLOVA, Korea's First Hyperscale 'Al to Empower Everyone'." *Naver Corp. Press Releases*, May 25, 2021. https://www.navercorp.com/en/promotion/pressReleasesView/30686.

Nightingale, Sophie J., and Hany Farid. "AI-synthesized faces are indistinguishable from real faces and more trustworthy." *PNAS* 119, no. 8 (February 2022). ISSN: 10916490. https://doi.org/10.1073/PNAS.2120481119.

Nimmo, Ben. *The Breakout Scale: Measuring the impact of influence operations*. Brookings Institution, September 2020. https://www.brookings.edu/research/the-breakout-scale-measuring-the-impact-of-influence-operations/.

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

"OPT-175B License Agreement." Metaseq. https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/MODEL_LICENSE.md.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. "Training language models to follow instructions with human feedback." *OpenAI*, March 2022. https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

Ovadya, Aviv. "'Contextualization Engines' can fight misinformation without censorship." *Medium*, May 26, 2022. https://aviv.medium.com/contextualization-engines-can-fight-misinformation-without-censorship-c5c47222a3b7.

Ovadya, Aviv, and Jess Whittlestone. "Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning." *arxiv:1907.11274 [cs.CY]*, July 2019. https://doi.org/10.48550/arxiv.1907.11274.

Pajola, Luca, and Mauro Conti. "Fall of Giants: How popular text-based MLaaS fall against a simple evasion attack." *Proceedings - 2021 IEEE European Symposium on Security and Privacy, Euro S and P 2021*, April 2021, 198–211. https://doi.org/10.48550/arxiv.2104.05996.

Papernot, Nicolas, Ian Goodfellow, Martín Abadi, Kunal Talwar, and Úlfar Erlingsson. "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data." *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings,* October 2016. https://doi.org/10.48550/arxiv.1610.05755.

Park, Hee Sun, Timothy R. Levine, Catherine Y.Kingsley Westerman, Tierney Orfgen, and Sarah Foregger. "The Effects of Argument Quality and Involvement Type on Attitude Formation and Attitude Change: A Test of Dual-Process and Social Judgment Predictions." *Human Communication Research* 33, no. 1 (January 2007): 81–102. ISSN: 0360-3989. https://doi.org/10.1111/J.1468-2958.2007.00290.X.

"Partnership on AI." Partnership on AI. Accessed October 29, 2022. https://partnershiponai.org/.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. "Shifting attention to accuracy can reduce misinformation online." *Nature* 592 (7855 2021): 590–595. ISSN: 1476-4687. https://doi.org/10.1038/s41586-021-03344-2.

Percy, Sarah. *Mercenaries: The History of a Norm in International Relations*. 1–280. Oxford University Press, October 2007. ISBN: 9780191706608.

"Public Comments to the Federal Communications Commission about Net Neutrality Contain Many Inaccuracies and Duplicates." *Pew Research Center,* November 29, 2017. https://www.pewresearch.org/internet/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

Pillars of Russia's Disinformation and Propaganda Ecosystem. U.S. Department of State, August 2020. https://www.state.gov/russias-pillars-of-disinformation-and-propaganda-report/.

"Poland: First GDPR fine triggers controversial discussions." ePrivacy Blog, May 17, 2019. https://blog.eprivacy.eu/?p=544.

"Prompt Engineering." co:here. https://docs.cohere.ai/docs/prompt-engineering.

Radford, Alex, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. "Better Language Models and Their Implications." OpenAI Blog, February 14, 2019. https://openai.com/blog/better-language-models/.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical Text-Conditional Image Generation with CLIP Latents." *arxiv:2204.06125 [cs.CV],* April 2022. https://doi.org/10.48550/arxiv.2204.06125.

Rashkin, Hannah, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. "Measuring Attribution in Natural Language Generation Models." *arxiv:2112.12870 [cs.CL],* August 2, 2022. https://doi.org/10.48550/arxiv.2112.12870.

Rawnsley, Adam. "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.

"Representative Anna Eshoo to Jake Sullivan and Alondra Nelson," September 20, 2020. https://eshoo.house.gov/sites/eshoo.house.gov/files/9.20.22LettertoNSCandOSTPonStabilityAI.pdf.

"Responsible AI Licenses (RAIL)." Responsible AI Licenses (RAIL). Accessed September 14, 2022. https://www.licenses.ai/.

Rid, Thomas. *Active Measures: The Secret History of Disinformation and Political Warfare*. 260. New York: Farrar, Straus, Giroux, 2020. https://us.macmillan.com/books/9780374287269/activemeasures.

Riedl, Martin J., Sharon Strover, Tiancheng Cao, Jaewon R. Choi, Brad Limov, and Mackenzie Schnell. "Reverse-engineering political protest: the Russian Internet Research Agency in the Heart of Texas." *Information, Communication, and Society* 25, no. 15 (2021). ISSN: 14684462. https://doi.org/10.1080/1369118X.2021.1934066.

Roozenbeek, Jon, Sander van der Linden, and Thomas Nygren. "Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures." *Harvard Kennedy School Misinformation Review* 1, no. 2 (February 2020). https://doi.org/10.37016//MR-2020-008.

Ruder, Sebastian. "Recent Advances in Language Model Fine-tuning." Sebastian Ruder (Blog), February 24, 2021. https://ruder.io/recent-advances-lm-fine-tuning/.

Sablayrolles, Alexandre, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. "Radioactive data: tracing through training." *37th International Conference on Machine Learning, ICML 2020* PartF168147-11 (February 3, 2020): 8296–8305. https://doi.org/10.48550/arxiv.2002.00937.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar, et al. "Imagen: Text-to-Image Diffusion Models." https://imagen.research.google/.

Sayler, Kelly M., and Laurie A. Harris. "Deep Fakes and National Security." *Congressional Research Services*, 2022. https://crsreports.congress.gov.

Schmidt, Eric, Robert Work, Safra Catz, Eric Horvitz, Steve Chien, Andrew Jassy, Mignon Clyburn, and et al. *Final Report*. National Security Commission on Artificial Intelligence, 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf#page=52.

Schneider, Jordan, and Irene Zhang. "New Chip Export Controls and the Sullivan Tech Doctrine with Kevin Wolf." ChinaTalk, October 11, 2022. https://www.chinatalk.media/p/new-chip-export-controls-explained.

Schneier, Bruce. "Toward an Information Operations Kill Chain." Lawfare, April 24, 2019. https://www.lawfareblog.com/toward-information-operations-kill-chain.

Schoen, Fletcher, and Christopher J. Lamb. *Deception, Disinformation, and Strategic Communications: How One Interagency Group Made a Major Difference*. Center for Strategic Research Institute for National Strategic Studies, June 2012. https://ndupress.ndu.edu/Portals/68/Documents/stratperspective/inss/Strategic-Perspectives-11.pdf.

Schwartz, Oscar. "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation." *IEEE Spectrum*, November 25, 2019. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

Sedova, Katerina, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan. *AI and the Future of Disinformation Campaigns: Part 2: A Threat Model*. Center for Security and Emerging Technology, December 2021. https://doi.org/10.51593/2021CA011.

Seger, Elizabeth, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó Heigeartaigh, and Helena Bacon. *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*. The Alan Turing Institute, October 14, 2020. https://doi.org/10.17863/CAM.64183.

Senate Report No 116-290, vol 2. 2020. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. "Compute Trends Across Three Eras of Machine Learning." *Proceedings of the International Joint Conference on Neural Networks,* March 9, 2022. https://doi.org/10.48550/arxiv.2202.05924.

Sevilla, Jaime, Pablo Villalobos, Juan Felipe Cerón, Lennart Heim Matthew Burtell, Amogh B. Nanjajjar, Anson Ho, Tamay Besiroglu, and Marius Hobbhahn. "Parameter, Compute and Data Trends in Machine Learning," 2021. https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=0.

Shannon, Vaughn P. "Norms Are What States Make of Them: The Political Psychology of Norm Violation." *International Studies Quarterly* 44, no. 2 (June 2000): 293–316. ISSN: 0020-8833. https://doi.org/10.1111/0020-8833.00159.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, January 2017. https://doi.org/10.48550/arxiv.1701.06538.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks against Machine Learning Models." *Proceedings - IEEE Symposium on Security and Privacy*, October 2016, 3–18. ISSN: 10816011. https://doi.org/10.48550/arxiv.1610.05820.

"So you're ready to get started." Common Crawl. Accessed June 27, 2022. https://commoncrawl.org/the-data/get-started/.

Solaiman, Irene, Miles Brundage, Openai Jack, Clark Openai, Amanda Askell Openai, Ariel Herbert-Voss, Jeff Wu Openai, et al. "Release Strategies and the Social Impacts of Language Models." *arxiv:1908.09203 [cs.CL],* August 2019. https://doi.org/10.48550/arxiv.1908.09203.

"Spot The Troll." Clemson University Media Forensics Hub. https://spotthetroll.org/.

Starbird, Kate, Ahmer Arif, and Tom Wilson. "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations." *Proceedings of the ACM on Human-Computer Interaction Vol: CSCW, Article 127,* CSCW 2019. ISSN: 25730142. https://doi.org/10.1145/3359229.

Strobelt, Hendrik, and Sebastian Gehrmann. "Catching a Unicorn with GLTR: A tool to detect automatically generated text." *gltr.io*. Accessed October 29, 2022. http://gltr.io/.

Stubbs, Jack. "Russian operation masqueraded as right-wing news site to target U.S. voters." Reuters, October 1, 2020. https://www.reuters.com/article/usa-election-russia-disinformation/exclusive-russian-operation-masqueraded-as-right-wing-news-site-to-target-u-s-voters-sources-idUSKBN26M5OP.

Stubbs, Jack, and Joseph Menn. "Facebook suspends disinformation network tied to staff of Brazil's Bolsonaro." *Reuters*, July 8, 2020. https://www.reuters.com/article/us-facebook-disinformation-brazil/facebook-suspends-disinformation-network-tied-to-staff-of-brazils-bolsonaro-idUSKBN2492Y5.

Sunstein, Cass R. "Social Norms and Social Roles." *Columbia Law Review* 44 (1996): 909. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=12456&context=journal_articles.

Sutskever, Ilya, James Martens, and Geoffrey Hinton. "Generating Text with Recurrent Neural Networks." Edited by Lisa Gooter and Tobias Scheffer. *Proceedings of the 28th International Conference on Machine Learning*, 2011. https://icml.cc/2011/papers/524_icmlpaper.pdf.

Tannenwald, Nina. "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use." *International Organization* 53, no. 3 (1999): 433–468. https://www.jstor.org/stable/2601286.

Taylor, Philip M. *Munitions of the mind: a history of propaganda from the ancient world to the present era*. Manchester University Press, 2003. ISBN: 978-1-84779-092-7.

Ternovski, John, Joshua Kalla, and Peter Aronow. "The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments." *Journal of Online Trust and Safety* 1, no. 2 (February 2022). ISSN: 2770-3142. https://doi.org/10.54501/JOTS.V1I2.28.

"The Internet Of Humans." Proof Of Humanity. Accessed October 29, 2022. https://www.proofofhumanity.id/.

The Weaponization of Information: The Need for Cognitive Security. RAND Corporation, April 27, 2017.

Thiel, David, Renee DiResta, Shelby Grossman, and Elena Cryst. *Contours and Controversies of Parler*. Stanford Internet Observatory, 2021. https://fsi.stanford.edu/publication/contours-and-controversies-parler.

Thiel, David, and Miles McCain. *Gabufacturing Dissent: An in-depth analysis of Gab*. Stanford Internet Observatory, 2022. https://cyber.fsi.stanford.edu/publication/gabufacturing-dissent-an-in-depth-analysis-of-gab.

"Ticks or it didn't happen." WITNESS Media Lab, December 2019. https://lab.witness.org/ticks-or-it-didnt-happen/.

Tiku, Nitasha. "The Google engineer who thinks the company's AI has come to life." *Washington Post*, June 11, 2022. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Tramer, Florian, Fan Zheng, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. "Stealing Machine Learning Models via Prediction APIs." *25th USENIX Security Symposium (Austin, TX; USENIX Security 16)*, 2016, 601–618. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer.

"Federal Rulemaking: Selected Agencies Should Fully Describe Public Comment Data and Their Limitations." *U.S. GAO*, September 2021. https://www.gao.gov/products/gao-21-103181.

U.S. Senate, Committee on Homeland Security, Permanent Subcommittee on Investigations, and Governmental Affairs. *Abuses of the Federal Notice-and-Comment Rulemaking Process*. 2019. https://tinyurl.com/5bamt57s.

United States, Supreme Court of the. "Van Buren v. United States," October 2020. https://www.supremecourt.gov/opinions/20pdf/19-783_k53l.pdf.

Venigalla, Abhinav, and Linden Li. "Mosaic LLMs (Part 2): GPT-3 quality for <$500k." Mosaic, September 29, 2022. https://www.mosaicml.com/blog/gpt-3-quality-for-500k.

Verdoliva, Luisa. "Media Forensics and DeepFakes: An Overview." *IEEE Journal on Selected Topics in Signal Processing* 14, no. 5 (January 2020): 910–932. ISSN: 19410484. https://doi.org/10.1109/JSTSP.2020.3002101.

"Verifiable Credentials Data Model v1.1." W3C, March 3, 2022. https://www.w3.org/TR/vc-data-model/.

Vincent, James. "YouTuber trains AI bot on 4chan's pile o' bile with entirely predictable results." *The Verge*, June 8, 2022. https://www.theverge.com/2022/6/8/23159465/youtuber-ai-bot-pol-gpt-4chan-yannic-kilcher-ethics.

Wallace, Eric, Tony Z. Zhao, Shi Feng, and Sameer Singh. "Concealed Data Poisoning Attacks on NLP Models." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021, 139–150. https://doi.org/10.48550/arxiv.2010.12563.

Walsh, Bryan. "OpenAI's GPT-3 gets a little bit more open." Axios, November 18, 2021. https://www.axios.com/2021/11/18/openai-gpt-3-waiting-list-api.

Wanless, Alicia, and James Pamment. "How Do You Define a Problem Like Influence?" *Journal of Information Warfare* 18, no. 3 (2019): 1–14. https://www.jstor.org/stable/26894679.

Wardle, Claire. "The Media Has Overcorrected on Foreign Influence." *Lawfare*, October 26, 2020. https://www.lawfareblog.com/media-has-overcorrected-foreign-influence.

———. "This Video May Not Be Real." *New York Times*, August 19, 2019. https://www.nytimes.com/2019/08/14/opinion/deepfakes-adele-disinformation.html.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. "Emergent Abilities of Large Language Models." *arxiv:2206.07682 [cs.CL]*, June 2022. https://doi.org/10.48550/arxiv.2206.07682.

Weng, Lilian. "Controllable Neural Text Generation." Lil'Log, January 2, 2021. https://lilianweng.github.io/posts/2021-01-02-controllable-text-generation/.

"Why you might be asked to upload a video selfie to confirm your identity on Instagram." Facebook Help Centre. Accessed October 29, 2022. https://m.facebook.com/help/1053588012132894.

Wiggers, Kyle. "Announcing AI21 Studio and Jurassic-1 Language Models." AI21 Labs. Accessed January 31, 2022. https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1.

———. "Huawei trained the Chinese-language equivalent of GPT-3." VentureBeat, April 29, 2021. https://venturebeat.com/ai/huawei-trained-the-chinese-language-equivalent-of-gpt-3/.

Woolley, Samuel C., and Douglas Guilbeault. "Computational propaganda in the United States of America: Manufacturing consensus online." *Project on Computational Propaganda Research*, 2017, 1–29.

Wu, Tongshuang, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. "PromptChainer: Chaining Large Language Model Prompts through Visual Programming." *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, April 2022. https://doi.org/10.1145/3491101.3519729.

Wu, Xingjiao, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. "A Survey of Human-in-the-loop for Machine Learning." *Future Generation Computer Systems* 135 (August 2021): 364–381. https://doi.org/10.1016/j.future.2022.05.014.

Xiang, Tao, Chunlong Xie, Shangwei Guo, Jiwei Li, and Tianwei Zhang. "Protecting Your NLG Models with Semantic and Robust Watermarks." *arxiv:2112.05428 [cs.MM]*, December 10, 2021. https://doi.org/10.48550/arxiv.2112.05428.

Yu, Jiahui, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, and Burcu Karagol Ayan. “Parti: Pathways Autoregressive Text-to-Image Model.” https://parti.research.google/.

Zeng, Wei, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, et al. “PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation.” *arxiv:2104.12369 [cs.CL]*, April 2021. https://doi.org/10.48550/arxiv.2104.12369.