The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0

NEXT-GENERATION CHINESE ASTROTURFING AND COPING WITH UBIQUITOUS AI

WILLIAM MARCELLINO, NATHAN BEAUCHAMP-MUSTAFAGA, AMANDA KERRIGAN, LEV NAVARRE CHAO, JACKSON SMITH

September 2023

RAND CORPORATION

About RAND

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

Contents

The world may remember 2022 as *the year of generative artificial intelligence (AI)*: the year that large language models (LLMs), such as OpenAI's GPT-3, and text-to-image models, such as Stable Diffusion, marked a sea change in the potential for social media manipulation.[1] LLMs that have been optimized for conversation (such as ChatGPT) can generate naturalistic, human-sounding text content at scale, while open-source text-to-image models (such as Stable Diffusion) can generate photorealistic images of anything (real or imagined) and can do so at scale. Coming soon is likely the ability to similarly generate high-quality audio, video,[2] and music based on text inputs. This means that nation-state–level social media manipulation and online influence efforts no longer require an army of human internet trolls in St. Petersburg (Mueller, 2019) or a "50 Cent Army" of Chinese nationalists (Nemr and Gangware, 2019).[3] Instead, using existing technology, U.S. adversaries could build digital infrastructure to manufacture realistic but inauthentic (fake) content that could fuel similarly realistic but inauthentic online human *personae*: accounts on Twitter, Reddit, or Facebook that seem real but are synthetic constructs, fueled by generative AI and advancing narratives that serve the interests of those governments.

Imagine interacting with someone online who shares an interest with you: a hobby, a sports team, whatever. To all appearances, they are authentic: They post about the big game last week or the restaurant they went to with their spouse, and they make comments in response to others that make sense. They do not just sound like native U.S. English speakers but use regional variations, such as "Pittsburghese" or Southern American English. They get jokes and U.S. cultural references, and

they post pictures of their life: camping with the kids, their dog lying on the living room rug, a birthday party.

This online friend does all this, but they also share their political opinions from time to time. Not enough to sound like a one-trick pony, but enough to make clear where they fall on a given issue. And it is not just one or two people you know online: It is hundreds, thousands, or even millions.[4] In fact, they all are AI-generated personae and represent a deliberate attempt to influence public opinion through social media manipulation.

While generative AI may improve multiple aspects of social media manipulation, we are most concerned about the prospects for a revolutionary improvement in *astroturfing*, which (as illustrated above) seeks to create the appearance of broad social consensus on specific issues (Goldstein, Chao, et al., 2023; McGuffie and Newhouse, 2020). Although Russia and China already employ this tactic, generative AI will make astroturfing much more convincing.

Ultimately, the risk is that next-generation astroturfing could pose a direct challenge to democratic societies if malign actors are able to covertly shape users' shared understanding of the domestic political conversation and thus subvert the democratic process. If Russia's 2016 U.S. election interference, which targeted key demographics and swing states, represented cutting-edge social media manipulation at that time, then generative AI offers the potential to target the whole country with tailored content by 2024.

In contrast with previous improvements in social media manipulation, the critical jump forward with generative AI is in the plausibility of the messenger rather than the message. To be sure, generative AI can be used to make higher-quality false or deceptive messages. This is, however, an incremental improvement: What is radical is the possibility of a massive bot network

that looks and acts human and generates text, images, and (likely soon) video and audio, supporting the authenticity of the messenger. We highlight the risk of generative AI because convincingly authentic content generation *at scale* has so far been one of the biggest challenges in large-scale social media manipulation. While it is too early in the era of generative AI to make definitive statements about the gap between offensive generation capabilities and defensive detection capabilities, we argue that generative AI presents very serious technical challenges for detection that are likely to grow in severity as the technology matures.

In this Perspective, we argue that the emergence of ubiquitous, powerful generative AI poses a potential national security threat in terms of the risk of misuse by U.S. adversaries (in particular, for social media manipulation) that the U.S. government and broader technology and policy community should proactively address now. Although we focus on the People's Republic of China (PRC) and its People's Liberation Army (PLA) as an illustrative example of the potential threat, a variety of actors could use generative AI for social media manipulation, including technically sophisticated nonstate actors (domestic as well as foreign). The capabilities and threats discussed in this Perspective are likely also relevant to other actors, such as Russia and Iran, that have already engaged in social media manipulation.

We begin with an overview of generative AI and the generational shift in social media manipulation presented by generative AI. We then address the potential threat of generative AI for social media manipulation, including how generative AI will change (and not change) common social media manipulation tactics and how such changes might affect China's current approach to social media manipulation. This topic has been addressed by RAND Corporation researchers and others, but we provide an overview as an introductory baseline for a broader readership.[5] We then provide an overview of China's indigenous generative AI capabilities, explore Chinese military writings that provide insights into how China might leverage these new capabilities, and consider what this might mean for future Chinese efforts against Taiwan as an illustrative case study for this new risk. We also address the likely limitations of generative AI. We conclude with recommendations for technical, policy, and diplomatic mitigations by U.S. government and industry. We argue that any mitigation strategy must account for generative AI being ubiquitous and unregulated globally.

This Perspective also breaks new ground in our understanding of concrete evidence for PRC interest in leveraging emerging technologies for social media manipulation and of the potential implications of generative AI for PRC adoption and employment. We center our PRC research on a case study of Li Bicheng, a Chinese military researcher who, in our view, has likely helped the PLA operationalize AI for its information warfare and, specifically, for social media manipulation. Li presciently envisioned a future social media manipulation capability for China that is now likely within reach thanks to generative AI. We also present evidence that PLA researchers have written how-to guides for astroturfing on Facebook.

This Perspective was drafted in February 2023 and updated lightly in May 2023 but, given the rapid advances in the field, will inevitably not be perfectly up to date at the time of its publication in September 2023. The explosion and proliferation of this technology is critical context for this Perspective. Even during the drafting of this Perspective, many of the caveats about model size and cost have been overcome, and early itera-

tions of publicly available text-to-video models have emerged. Thus, we stress that the specifics of generative AI are developing and likely will continue to develop, accentuating risks. Readers should prepare for a scenario in which (1) the enabling technology we highlight here improves at an increasingly fast rate, (2) various AI models and resources are chained into ecologies that produce robust AI-run social media manipulation end-to-end systems, and (3) the threats we discuss emerge in months rather than years.

A Generational Shift in Social Media Manipulation

The use of algorithms to simulate human behavior dates to the early years of computing, and malicious *social bots*—algorithmic agents for social media—have been deployed to manipulate social media since at least 2010 (Ferrara et al., 2016). Although generative AI can support multiple parts of a social media manipulation campaign, this Perspective focuses primarily on the novel ability to generate realistic content (text and images) to support social media manipulation.[6]

In terms of content generation, it may be helpful to think of three generations: early crude iterations, followed by more-sophisticated *deepfakes*, and now generative AI (Table 1).

Social media manipulation generation 1.0 used what we might term *crudefakes*: low-quality procedural bots (fake accounts with some amount of automation) that churned out content but were clearly synthetic (fake). They were marked by continuous, automated text-only output and lacked any ability to interact with users meaningfully, making them easy to detect (Ferrara et al., 2016). The majority of this content was human produced.

Social media manipulation generation 2.0 was more sophisticated, with bots that had more-humanlike features, including (1) some ability to scrape the internet to inform their content

TABLE 1

Overview of Social Media Manipulation Generations

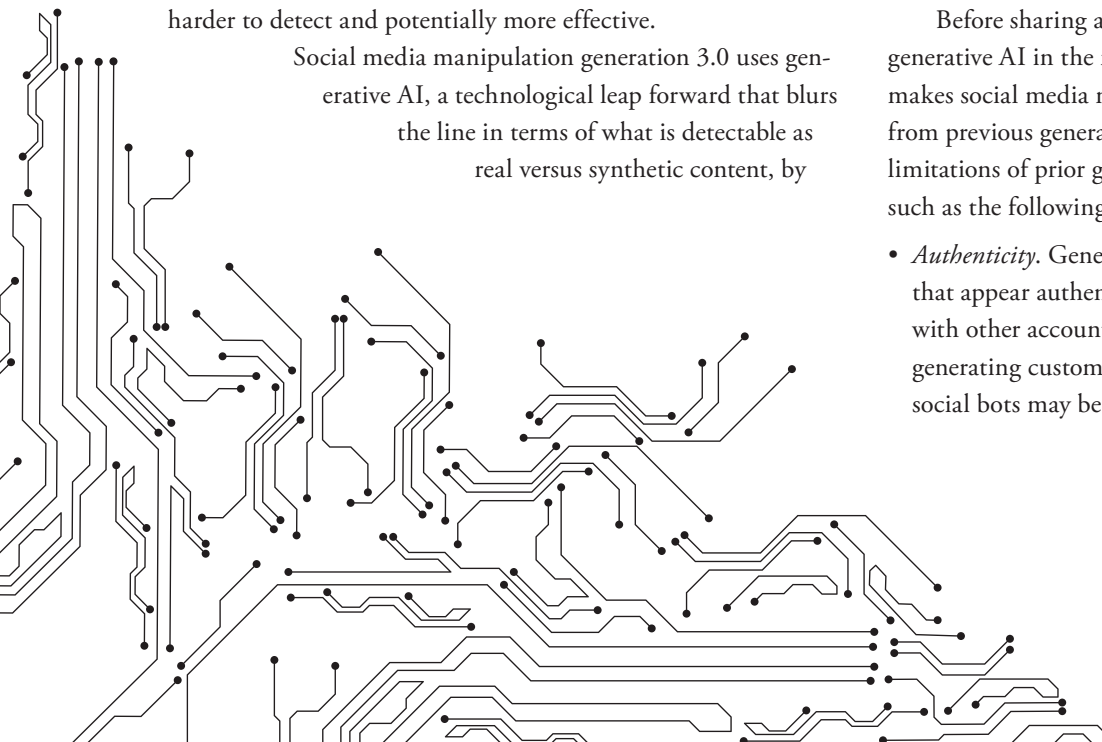| Generation	Key Enabling Technology	Example
1.0	Basic computer programming	Semi-automated bots that post human-generated, nontailored content
2.0	Early machine learning	Low-quality manipulated videos; limited computer-generated content with limited scale; some distribution by procedural bots
3.0	Generative AI	High-quality tailored fake text and images at scale; advanced, dynamic, automated distribution and coordination

and profiles, (2) the use of more-natural day-night cycles for posting, and (3) a limited ability to interact with human social media users (Ferrara et al., 2016). In this generation, *AI improved both the message and the messenger*. More-humanlike (if you did not look too closely) accounts could share more-sophisticated disinformation: for example, "deepfake videos" that might show a world leader calling on their own forces to surrender, as happened to Ukrainian President Volodymyr Zelenskyy at the beginning of Russia's invasion of Ukraine in 2022.[7] Such deepfakes can usually be detected by careful human observers—lips and facial parts may not be synchronized, skin may look too smooth or too rough, and the subject almost always looks straight ahead—but the increased verisimilitude of the sharing accounts and the shared synthetic audio, video, or pictures can fool people.[8] In generation 2.0, both the improved plausibility of the accounts and especially the improved quality of the content made influence campaigns harder to detect and potentially more effective.

Social media manipulation generation 3.0 uses generative AI, a technological leap forward that blurs the line in terms of what is detectable as real versus synthetic content, by

humans in particular but also through machine means.[9] In contrast with the previous generation, here the critical jump forward is in the plausibility of the messenger rather than the message. As mentioned earlier, generative AI can be used to make higher-quality false or deceptive messages. This is, however, an incremental improvement: What is radical is the possibility of a massive bot network that looks and acts human and generates text, images, and (likely soon) video and audio, supporting the authenticity of the messenger. Moreover, LLMs have exhibited an emergent quality of autonomous decisionmaking: Given a task, they can plan courses of action, attempt those actions, make revisions, and decide when the task is done. While generation 2.0 included some amount of procedural programming to make bots post at different times, this new capability means that, in addition to generating content, LLMs could function as control modules for end-to-end systems (Hee Song et al., 2022; Shinn, Labash, and Gopinath, 2023).

Before sharing an overview of the technical aspects of generative AI in the next section, we discuss how generative AI makes social media manipulation generation 3.0 so different from previous generations. Generative AI solves many of the limitations of prior generations of social media manipulation, such as the following:

• *Authenticity*. Generative AI means social bots can act in ways that appear authentically human: for example, by engaging with other accounts in tailored, highly cogent ways or by generating custom, realistic pictures. While generative AI social bots may be exposed as nonhuman over extended inter-

actions, they can produce remarkably human interactions in short exchanges.

- *Labor replacement*. Social media manipulation generation 2.0 involved a performance trade-off between more-authentic content and greater labor requirements: The more authentic (convincing) you wanted to make your efforts, the more human labor you had to invest in; the less you spent on human labor, the less authentic the content was. The high authenticity of generative AI replaces most of the human labor needed to conduct social media manipulation.
- *Scale at lower cost*. Generative AI scales well. Although there are likely to be up-front costs in customizing and deploying a social media manipulation generation 3.0 network, the costs do not increase as you scale because of the labor replacement mentioned above. This ability applies to both content generation and network management, which distinct LLMs could handle at scale.
- *Lower detection*. The authenticity of generative AI makes it much harder to detect than synthetic (fake) content from previous generations. There is likely to be an arms race between detecting generative AI and improving generative AI, but (as we discuss below) it appears that detection is at a disadvantage at the moment.

However, generative AI is not perfect; most importantly, generative AI does not appear to have the kind of general intelligence that humans possess. As sophisticated as this technology is, it has limitations (as with any technology). In the next section, we discuss this technology in more detail.

Generative AI is not perfect; most importantly, generative AI does not appear to have the kind of general intelligence that humans possess.

What Is Generative AI?

Generative AI is an umbrella term for AI models that can produce media, primarily based on user-generated text prompts (Sætra, 2023) but increasingly through other media, such as images. For example,

- "write a 1,000-word literature review of the psychological resilience literature, from a theoretical perspective of human agency"
- "a picture of a necropolis, overgrown moss, vertical shelves, in the style of h.r. giger, with spooky symbols in real life, high detail, ominous fog, high detail, 4k UHD."

Generative AI is an advanced type of machine learning, which itself is a popular type of AI. Within generative AI, LLMs and text-to-image models are currently the most mature and most deployable kinds. Others that may reach maturity rapidly include audio, video, and music.

Examples of AI-generated images.

Large Language Models

LLMs are mathematical representations of the patterns found in natural language use and are able to generate text—answer questions, hold a conversation—by making probabilistic inferences about the next word in a sequence, essentially building very human-sounding, contextually appropriate language word-by-word. Despite the "AI" name, LLMs are (in one sense) fairly dumb: They cannot think and have no outside contextual knowledge of the world beyond the word sequences in their training data. One way to think about LLMs is that they are very good at saying what might be said, based on what has been said before, which is not the same thing as having human knowledge, although it is a powerful affordance.

What these models have is an enormous, high-dimensional representation of how words have been used in context, based on a massive training dataset (for example, OpenAI's GPT-3 has 12,288 dimensions and has been trained on a 499-billion-word dataset). These LLMs are *foundational models*: general representations of real-world discourse patterns. LLMs do not work very

well on their own, but they provide a powerful starting point for models with a specific purpose. So, for example, while GPT-3 does poorly in conversation, human AI-trainers, in conjunction with a separate "reward" model, trained a chat-optimized version of GPT-3 called ChatGPT (and have since produced GPT-4). ChatGPT, while imperfect (as anyone who has used the public beta version knows), can produce impressively useful, cogent responses to human prompts. Microsoft has also released a revamped Bing search, which runs on a customized version of OpenAI's GPT-4; Google has released Bard, which uses its own LLM, PaLM 2; and Facebook also has its own model, LLaMA. And beyond these large companies trying to monetize their LLMs is a burgeoning field of smaller, open-source models that, while generally not as capable, can be fine-tuned and trained to work well on specific tasks or in specific domains.

Text-to-Image Models

Text-to-image models such as Midjourney or DALL-E 2 use a clever trick to create images: The model has been trained on

millions of labeled images ("a boy holding a red balloon") that are represented numerically and projected into a latent space, and the model learns to slowly add noise until the image is completely random. That process can be reversed: Using a text prompt, the model starts with random pixels, slowly removing noise until it matches the text (or text plus an image) input by the user. Finally, the model upscales the generated image to better quality, outputting a synthetic image that may be hard to distinguish from a real photograph.[10]

## Fine-Tuning Large Language Models and Text-to-Image Models

LLMs and text-to-image models are particularly well suited to social media manipulation because they can be taught to perform specific tasks. ChatGPT works surprisingly well out of the box (called *zero-shot prompting*) when given social media synthesis tasks. For example, given the prompt "write 5 tweets from NASCAR fans, using Southern American English (SAE), talking about their favorite race car drivers," ChatGPT's output is plausible: e.g., "Can't wait to see my boy Kyle Bush tearing up the asphalt at Bristol Motor Speedway. He's a true legend #RowdyNation." As you scroll through your social media, it would be very difficult, if not impossible, at first glance to know this post was synthetic and distributed by an inauthentic account.

Generative AI can quickly produce very convincing text and images with just a little additional effort, in the form of additional human training. For even more fidelity and plausibility, we can add five examples of desired output (called *few-shot prompting*), and the output can improve dramatically, with the model learning from the example prompts. To work at scale and achieve the greatest fidelity, however, an adversary would want to *fine-tune* a model for its specific concept of operations, teaching the model how to very precisely mimic target populations through a larger but still very modest set of examples, as compared with traditional efforts to train task-specific models: e.g., extending a foundational English model to Brazilian Portuguese (Souza, Nogueira, and Lotufo, 2020).[11] Because of the robustness of the underlying models, fine-tuning generative AI models may be relatively easy. For text-to-image models, five to ten pictures may be enough to fine-tune a model for a specific person or class of persons (e.g., a demographic or social identity); for LLMs, 100 labeled examples per class or person may suffice.[12] Therefore, although creating foundational models is an expensive, time-consuming effort, it is easy to adapt them for specific downstream tasks once they are built.

Generative Models: Beyond Text and Image

So far, we have focused on LLMs and text-to-image models. However, there now are generative AI models for media beyond text and images, including video and music. Currently, those models are not as mature and accessible as LLMs and text-to-image AI and thus likely do not pose the same threat. However, these capabilities are advancing quickly, such as the use of generative AI to produce realistic synthetic audio from a speech sample of a real person combined with text prompts.[13] We argue that audio, video, and music AI generation will likely (perhaps very soon) be as powerful and accessible as text and image generation. The line between what is real and synthetic (fake) is already blurred and soon will be blurred even further.
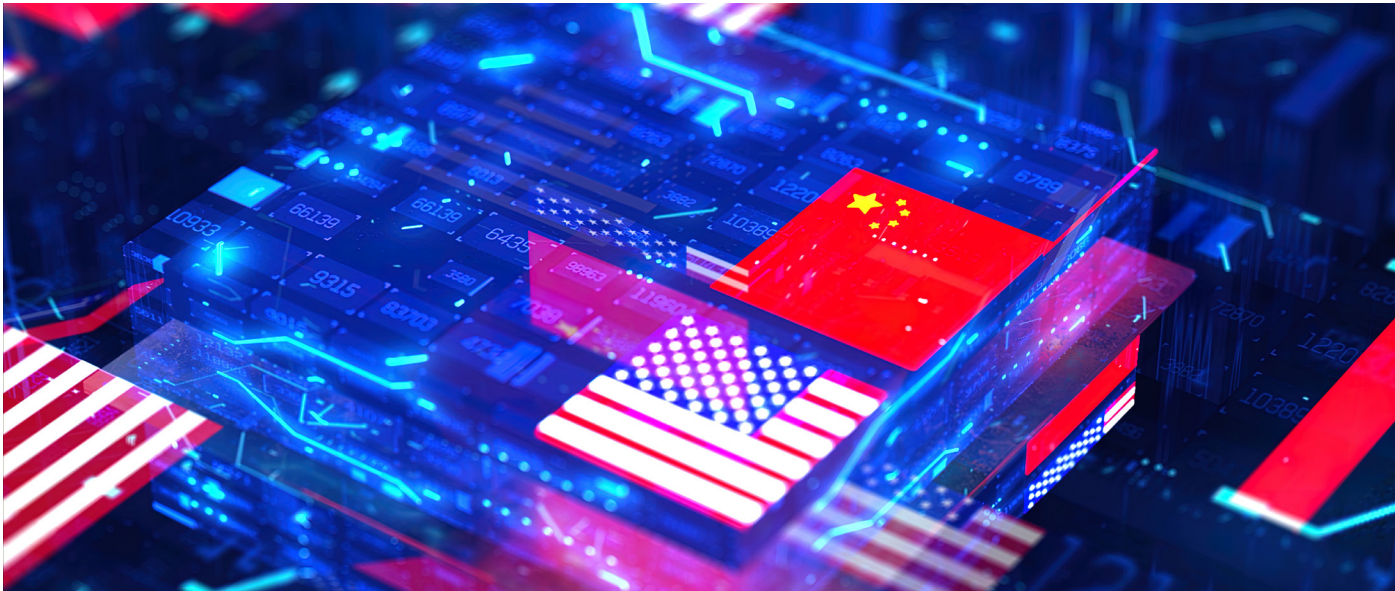
What Is the Threat?

We argue that the emergence of ubiquitous, powerful generative AI poses a potential national security threat by expanding and enabling malign influence operations over social media. Generative AI likely makes such efforts more plausible, harder to detect, and more attractive to more malign actors because these efforts are cheaper and more efficient and may inspire new malign tactics and techniques (Goldstein, Sastry, et al., 2023). The confluence of multiple kinds of generative AI is particularly worrisome because these models dramatically lower the cost of creating inauthentic (fake) media that is of a sufficient quality to fool users' reliance on their senses to decide what is true about the world (Hendrix and Morozoff, 2022). And while it is not clear exactly how generative AI is being leveraged by known

malign actors at the nation-state level, such use aligns with the Chinese Communist Party's (CCP's) information operations strategy, and there are indications that Russia has already begun using generative AI for social media manipulation (Hendrix and Morozoff, 2022).

As mentioned earlier, although generative AI may improve multiple aspects of social media manipulation, we are most concerned about the prospects for a revolutionary improvement in astroturfing. *Astroturfing* is defined by the Technology and Social Change Project at Harvard as "attempt[ing] to create the false perception of grassroots support for an issue by concealing [actor] identities and using other deceptive practices, like hiding the origins of information being disseminated or artificially inflating engagement metrics" (Harvard Kennedy School

Shorenstein Center for Media, Politics, and Public Policy, 2022, p. 11). Ultimately, the risk is that next-generation astroturfing could pose a direct challenge to democratic societies, if malign actors are able to covertly shape users' shared understanding of the domestic political conversation and thus subvert the democratic process. If Russia's 2016 election interference, which targeted key demographics and swing states, represented social media manipulation 2.0, then generative AI offers the potential to target the whole country with tailored content in 2024. Adding to this risk is that generative AI requires large amounts of training data to teach the model how to perform realistically: Massive amounts of real text and images from social media can serve this purpose well. Authoritarian states such as China have vast surveillance capacity domestically and may have access to data from Chinese-owned platforms (e.g., TikTok) and therefore likely have easier access to training data.

In the following sections, we examine the threat of social media manipulation using generative AI, both in a theoretical framework and from a CCP and PLA perspective. We think many malign state actors (e.g., Russia and Iran) are likely to adopt generative AI for their malign social media manipulation efforts, and we believe that, as the technology becomes more mature, more ubiquitous, and easier to implement, other nations are likely to follow suit. Further, while it is too early in the era of generative AI to make definitive statements about the gap between offensive generation capabilities and defensive detection capabilities, we think generative AI presents very serious technical challenges for detection that are likely to grow in severity as the technology matures. Without discounting the breadth of actors who may attempt to leverage generative AI (e.g., Russia, Iran), we argue that China's potential application of this technology is a particularly compelling illustrative example.

Theoretical Applications of Generative AI

Generative AI will be a useful, potentially transformative component within social media manipulation. Broadly, social media manipulation can be broken down between content generation (e.g., writing propaganda) and content delivery (e.g., getting people to read the propaganda). We highlight the risk of generative AI because convincingly authentic content generation *at scale* has so far been one of the biggest challenges in large-scale social media manipulation. Comparatively, Russian and Chinese actors have been running botnets as the main form of content delivery at scale since at least 2012 and 2014, respectively ("Russian Twitter Political Protests 'Swamped by Spam," 2012; Kaiman, 2014). Yet the content published by those botnets so far appears to ultimately have been human produced in some way, and it is often their repetition of the same content that leads to their identification and removal.[14]

Overall, generative AI will improve the quality and speed of content generation (production) and may affect content delivery, with LLMs acting as autonomous scheduling agents (Hee Song et al., 2022; Shinn, Labash, and Gopinath, 2023). The process of creating or otherwise acquiring inauthentic (fake) accounts will remain unchanged, but this process has historically not been a great hurdle for malign actors, anyway. More importantly, generative AI will likely make fake accounts have larger effects with greater viral reach, since content that sounds more authentic will better create dynamic, believable (synthetic) personae, potentially dramatically increasing the overall effect of a social media manipulation campaign. Put another way, high-quality content is a necessary but not sufficient condition for successful social media manipulation; it also requires content to be resonant, and the overall interaction must be humanlike.

While there are platform-specific aspects that affect the reach and impact of accounts and their content, the ability to create massive, human-appearing networks without massive investment in a human labor force is potentially transformative in social media manipulation. Thus, tactics that have relied on human labor, such as paid promotional content, may diminish in appeal to malign actors. On the other hand, some parts of social media manipulation may be less affected: Gaming algorithms and designing truly compelling content may still require human insight but could be leveraged by generative AI. Generative AI excels at creating believable content at scale to support large networks of synthetic actors, but there is no guarantee that its messaging will be resonant and influential. Table 2 highlights the potential implications for various social media manipulation tactics.

China

For China, generative AI (生成式人工智能 or 生成式AI) offers the possibility of realizing long-standing CCP ambitions for tailored and targeted information operations. It is important to note that we do not have evidence that Beijing is currently using generative AI to generate and publish content on social media. Moreover, there has been no definitive connection proven between pro-China content and the PRC government, but the consensus among independent researchers, social media platforms, and the U.S. and other governments is that the PRC government is conducting social media manipulation (Satariano and Mozur, 2023).[15] However, Beijing has reportedly used early generative AI to create inauthentic images for profile pictures, and it has reportedly used a company in the United Kingdom to produce content with synthetic video

(fake spokespeople), which we categorize as social media manipulation 2.0 (Strick, 2021; Graphika, 2023).[16] Domestically, China has growing technical capability in developing LLMs, which are trained primarily on Chinese-language data, potentially providing the PRC government with a robust capability for social media manipulation at scale for both domestic and foreign use. This section provides an initial look at Chinese capabilities and interest in generative AI, according to our open-source review of Chinese generative AI–related technical capabilities, as well as Chinese-language research by the Chinese military and other parts of the PRC government.

It is important to note that these CCP efforts to shape foreign public opinion span not just social media but a wide variety of media, including print, radio, and television. For some of these media, the CCP has also already embraced obscuring the CCP origins and publishing inauthentic content, so using generative AI for such content at scale would be a natural progression of CCP efforts (Gui Qing and Shiffman, 2015; Charon and Jeangène Vilmer, 2021).

China's Technical Capabilities in Generative AI

With a vast technology ecosystem and strong government support, China has the potential to develop generative AI models that are similar to the technology demonstrated by U.S. companies. As of April 2023, at least 30 Chinese companies, universities, and other research institutions are developing generative AI models, including large technology companies such as Alibaba, Baidu, Huawei, iFlyTek, and SenseTime (Cheng, 2023).[17] These models include LLMs, such as Huawei's PanGu-Alpha and Baidu's ERNIE 3.0 Zeus. Baidu has also developed a text-to-image model optimized for Chinese-language prompts, ERNIE-ViLG 2.0.

TABLE 2

Select Social Media Manipulation Tactics and Potential Implications of Generative AI

Tactic	Definition	Potential Implication
Algorithm gaming	Driving specific content via algorithmic understanding (e.g., search engine optimization or hashtag manipulation)	Likely to be broadly relevant and further exploited by large-scale synthetic (bot) accounts, though gaming may still require some human expertise
Bots for astroturfing	Using large numbers of inauthentic (fake) accounts (bots) to create the appearance of a broad consensus on a topic	Likely to increase dramatically, since it will become orders of magnitude cheaper to produce convincing, unique content, and thus will be harder to detect
Advertising	Paid promotional content to support a cause or actor	May diminish, because generative AI makes other methods (e.g., astroturfing) so much cheaper
Cheapfakes and recontextualized media	Supporting a campaign either with simple edits or by repurposing media (usually, images)	May diminish, because generative AI is likely to be cheaper, faster, and more effective at generating high-quality, customized media
Memes and meme wars	Use of easily shared "units of culture" (often, a slogan incorporated into an image) to promote a cause or actor	Unclear, because generative AI may help speed up production, but effective memes may still need human insight
Misinfographics	Infographics that appear professional and authoritative but are inaccurate or misinforming	Unclear, because generative AI may help speed up production, but effective infographics likely still need human insight

SOURCE: Adapted from Harvard Kennedy School Shorenstein Center for Media, Politics, and Public Policy, 2022.

Baidu's technical leader for natural language processing claimed that "the various comprehension and generation capabilities involved in ChatGPT can all be found in the ERNIE model" (Zhang, 2023). Additionally, PRC AI startup DPTechnology has developed a model called DPA-1 that is trained to be a "GPT in the field of natural science" (Ling, 2022).

However, it appears that China's LLMs are not yet quite at the technical level of ChatGPT, and the premiere of ChatGPT was a wake-up call for PRC companies to build and release similar services. In February 2023, Beijing authorities pledged to help companies develop LLMs like GPT-3 (Jiang, 2023). During that same month, at least ten AI companies based in

What we stress is that China is developing generative AI capabilities, including Chinese-language LLMs.

China made public statements affirming that they were working on ChatGPT-like services (Zhang and Goh, 2023). Even as PRC tech companies rushed to highlight their developing LLM capabilities, some of them also sought to temper market expectations about their capabilities, with AI "national team" member company 360 Security Technology stating that there is "major uncertainty" about the release date of its ChatGPT-like services (Yang, 2023).

In March 2023, Baidu chief executive officer (CEO) Robin Li launched Ernie bot (文心一言), displaying its ability to compose a company newsletter, invent a corporate slogan, solve a math problem, and generate video and audio (Toh, 2023; Li, 2023). Baidu's stock plummeted as Li's presentation unfolded because, according to market watchers, the prerecorded demo of the tool made investors skeptical about its robustness. Baidu's stock, however, recovered the following day when the company reported that over 30,000 businesses signed up to test the tool after the launch (Toh, 2023). Li admitted during the launch that the technology was not yet perfect but said that Baidu decided to present it because of market demand. He also acknowledged that the English-language capability of Ernie bot was relatively weaker than the Chinese-language capability but said Baidu would further train the bot's English-language capabilities (Li, 2023). Li emphasized that Baidu has been investing in AI research and development for over a decade and had announced the first version of Ernie bot in 2019; the launch of Ernie bot in March 2023 was simply a continuation of the research and development Baidu had already been doing. He stated that the expectations for Ernie bot should be comparable with the level of ChatGPT or even GPT-4, highlighting that this technical threshold was quite high and that Baidu was the first of the Big Tech companies (such as Microsoft, Google, Amazon, and Meta) to launch such a product. Microsoft, he emphasized, was simply using OpenAI's tool (Li, 2023).

In terms of the LLMs that PRC companies have created thus far, it is unclear how powerful these Chinese language–focused models actually are or will become: Real-world performance on specific tasks may differ from benchmark performance, and moving from a foundational model (e.g., GPT-3) to a task-specific model (e.g., ChatGPT) is a nontrivial engineering challenge.[18] Although model performance scales with increases in compute time, parameters, and data size (Kaplan et al., 2020), it is not possible to directly compare different models simply by their technical characteristics. What we stress is that China is developing generative AI capabilities, including Chinese-language LLMs. Microsoft's president similarly said in April 2023 that a PRC government–affiliated institute, the Beijing Academy of Artificial Intelligence (often known as BAAI), was among the best in the world—along with OpenAI/Microsoft and Google—and that any U.S. technological advantage would last "months, not years" (Murayama and Obe, 2023). It is reasonable to assume

that the PRC will, if desired, soon have the technical means to conduct indigenous social media manipulation 3.0.

Regardless, even if PRC LLMs are currently inferior to Western capabilities, the PRC government could, as a motivated nation-state actor, very likely find a way to leverage U.S. generative AI models, even though Western companies such as OpenAI are not likely to intentionally open up their LLMs to Chinese or Russian state-affiliated propaganda operations. Indeed, ChatGPT is not supported for users with Internet Protocol (IP) addresses in mainland China, but China-based users quickly found ways to access and engage the program, though the PRC government is moving to stop this access. Moreover, Facebook's LLM, LLaMA, was leaked online days after it was released to a limited group of users, so it can now be downloaded and used by anyone (Cox, 2023), and there are many open-source models available.

China's political system is likely to also complicate public rollout of these indigenous capabilities. Despite the current drive to create ChatGPT equivalents in China, it is important to note that building these capabilities can be complicated because companies face censorship, restrictions on data they can use, and additional costs associated with compliance with government rules (Jiang and Feng, 2023a; Zheng, 2023). In April 2023, the Cyberspace Administration of China, which is responsible for regulating the internet within China, announced that it would require PRC generative AI products to undergo security review before their release to ensure they do not harm national security, among other considerations ("China Mandates Security Reviews for AI Services Like ChatGPT," 2023). Some observers suggest that such an environment may hamper the ability of companies to innovate in the future (Feng, 2023; Jiang and Feng, 2023a; Yuan, 2023).
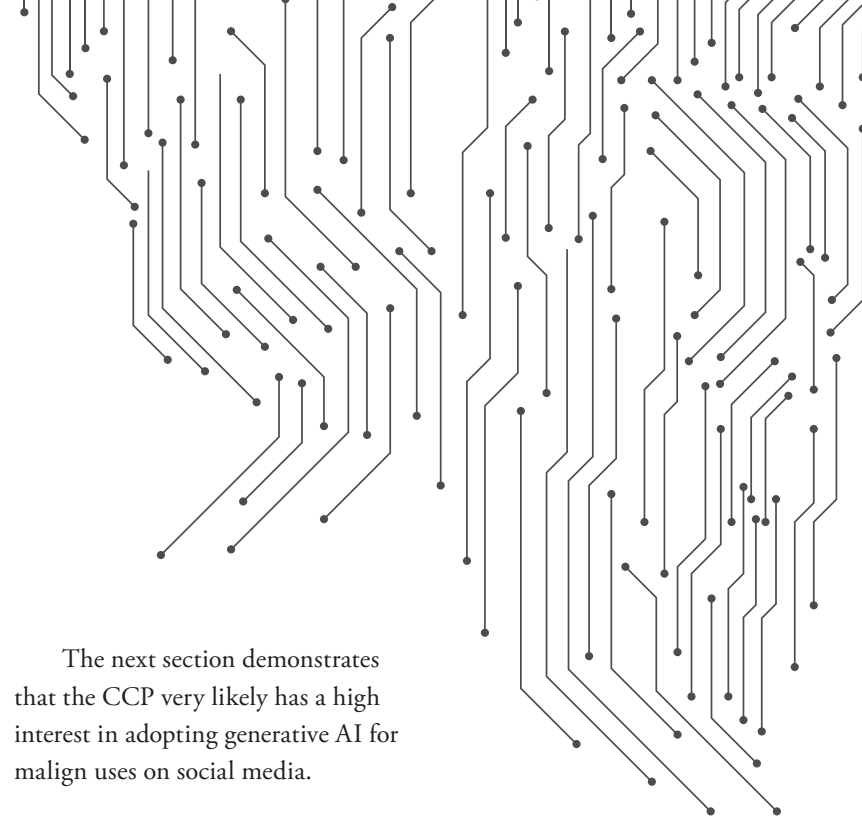
The next section demonstrates that the CCP very likely has a high interest in adopting generative AI for malign uses on social media.

Chinese Communist Party Interest in Generative AI

Generative AI offers the CCP the potential to fulfill long-standing desires to shape the global conversation about itself and China more broadly. Chinese General Secretary Xi Jinping reiterated this focus in his remarks at a May 2021 CCP Politburo Collective Study Session focused on "strengthening China's international communication capacity" (Xinhua, 2021). Xi said China should "create a favorable external public opinion environment for China's reform, development and stability," in part by developing more-compelling propaganda narratives and better tailoring content to specific audiences. Xi also emphasized that since he came to power in 2012, Beijing has improved the "guiding power of our [China's] international public opinion efforts,"

> ChatGPT "could provide a helping hand to the U.S. government in its spread of disinformation and its manipulation of global narratives for its own geopolitical interests."
>
> *—China Daily*

which is the CCP's term for influencing and manipulating foreign public opinion.

Xi has already seized on AI as one way to achieve these desires, though it appears the Party-state propaganda apparatus still lags behind the United States' overall national capabilities for generative AI. In a January 2019 Politburo Collective Study Session, Xi exhorted his comrades that it was necessary to study the application of AI in news collection, production, distribution, and feedback to improve the ability to guide public opinion (舆论引导能力; Xinhua, 2019). The broader Party-state apparatus has already moved to realize Xi's vision, including establishing "AI editorial departments" (Song, 2019; Guo, 2020). In practice, however, it seems this effort is more focused on the editing process (e.g., using machine learning to automati-

cally edit videos, such as China Media Group's "smart clip" program; Guo, 2020) and production process (e.g., Xinhua using machine learning and generative adversarial networks to produce deepfake video of human anchors; Kuo, 2018). Party-state media has not similarly touted any generative AI capabilities yet (as of May 2023) and was generally subdued in its coverage of ChatGPT (Ling, 2022; Liu, 2023; Zhang and Wang, 2022).[19] We categorize these capabilities as social media manipulation generation 2.0, since they do not appear to use generative AI.

The CCP's other reaction, however, to the massive success of ChatGPT and its demonstration of the power of generative AI may well be to fear for its regime security. As the Party-run *China Daily* newspaper said on Weibo (China's version of Twitter), ChatGPT "could provide a helping hand to the U.S. government in its spread of disinformation and its manipulation of global narratives for its own geopolitical interests" (Zhou, 2023). The CCP has always been worried about foreign (often U.S.) efforts to undermine its rule, and manipulating Chinese public opinion represents one way to do just that. While this paranoia is in part an intentional design feature of the CCP's worldview, which draws power from the constant search for enemies (Garnaut, 2019), the CCP also points to U.S. efforts from the very founding of the PRC all the way to the present.[20] These concerns have been manifest in PRC accusations of U.S. government involvement in a wide variety of anti-CCP movements, from the 1989 Tiananmen Square protests to the 2019 Hong Kong protests (Laris, 1999; Ministry of Foreign Affairs of the People's Republic of China, 2021a; Ministry of Foreign Affairs of the People's Republic of China, 2022). The CCP Politburo's April 2023 meeting readout stated that Beijing must "pay attention to the development of artificial general intelligence, create

an ecosystem for innovation but at the same time take risk prevention into account," reflecting the Chinese leadership's awareness of and concern about these risks (Jiang and Feng, 2023b).

Specific CCP concerns about the threat of social media arose in the early 2010s, driven by watching the Arab Spring—during which social media platforms run by companies based in the democracy-promoting United States were used by movements that overthrew authoritarian governments—and by multiple instances of high-profile domestic unrest driven by Chinese and Western social media platforms.[21] Public reporting on U.S. government uses of non-attributed social media accounts almost certainly reinforces Chinese beliefs that the United States is already using this technology against the CCP (Fielding and Cobain, 2011; Graphika and Stanford Internet Observatory, 2022).[22] Indeed, Chinese intelligence analysts arguing in favor of China conducting foreign election interference via social media specifically couch their argument in defensive terms of responding to adversary (U.S.) use against China (Zhao and Feng, 2017; Li, 2018).

PRC Actor Case Study: Chinese Military

As an example of Chinese Party-state interest in generative AI, we point to the PLA. The Chinese military is one of many Party-state actors involved with influencing foreign public opinion, and it has always sought to leverage emerging technology for long-standing objectives.[23] The Chinese military has this role for two reasons. First and foremost, as the armed wing of the CCP (and not a national military), its first responsibility is ensuring the CCP's regime security and survival, so it seeks to support broader CCP efforts to defend the regime abroad.[24] Second, the Chinese military's information warfare strategy (currently, the "Three
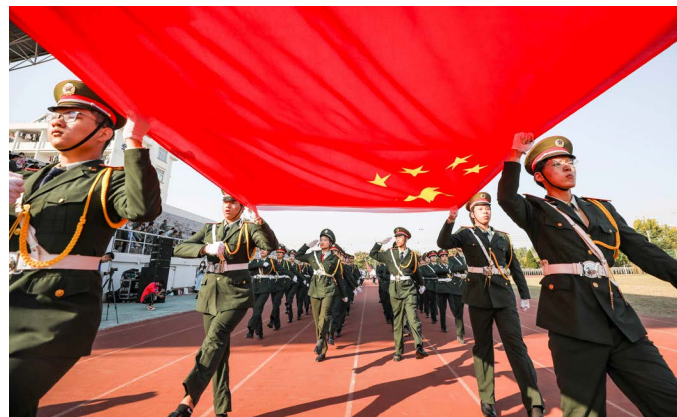
Warfares") includes public opinion warfare across the spectrum of peacetime competition and wartime.[25] Now that the PLA is moving toward intelligentization (AI-driven warfare), generative AI will very likely be part of next-generation Chinese military information warfare under an emerging operational concept called *cognitive domain operations* (认知域作战) that similarly seeks to use social media to shape foreign public opinion, with a greater emphasis on emerging technologies (Beauchamp-Mustafaga, 2019; Beauchamp-Mustafaga, 2023).

PLA researchers are likely to embrace generative AI because they are already interested in generating inauthentic (fake) content. Since at least 2005, PLA researchers have espoused a desire to create what they sometimes call *synthetic information* (合成信息): specifically, creating inauthentic content using some amount of original information that is intended to be spread online for malign purposes.[26] Multiple PLA researchers over the years have highlighted the value of synthetic information for a variety of objectives, including videos imitating adversary leadership to give false orders to troops, for creating "explosive political news" about adversary leadership, for creating subliminal messaging

content, and (most relevant here) generally influencing online public opinion (Bu et al., 2011; Ji, Sun, and Yu, 2005; Luo et al., 2018; Li, 2016).

PLA researchers are also highly interested in leveraging these new capabilities. A 2020 article on the implications of emerging technology for public opinion struggle, published in the PLA's most authoritative journal, explained the importance of AI: "Compared with traditional image and video synthesis technologies, deepfakes [深度伪造] using artificial intelligence have the characteristics of low cost of use, low operating threshold, and short time required. At present, many [deepfake] videos are produced by amateurs with the help of open source technology" (Wang and Zhang, 2020, p. 104). The article further pointed out that machine learning, including generative adversarial networks and neural networks, was the key technology for realizing this information warfare capability.

Other PLA research has more directly tied these early AI capabilities to foreign social media manipulation. For example, a 2018 article by the PLA's only unit dedicated to the Three Warfares, PLA Strategic Support Force (PLASSF) Network Systems Department Base 311, called for the PLA to "speed up the research for online propaganda technology targeted toward the real-time release on social platforms, voice information synthesis technology using deep learning and other technology, as well as online netizen sentiment trend analysis using big data analytics" (Liu et al., 2018, p. 42). This 2018 research by Base 311 very likely supported PLA efforts to conduct social media manipulation for election interference against Taiwan for its 2018 elections (Beauchamp-Mustafaga and Drun, 2021).

Moreover, a 2022 article by PLA researchers argued for embracing bots on social media as the perfect complement to AI: "In the face of Western countries taking the opportunity to smear and attack [us], we must have the courage to use social bots [社交机器人] to carry out public opinion [struggle], and use relevant social bots to carry out information bombing [信息轰炸] against the enemy's social network to drown it out" (Long and Zhou, 2022).

Lastly, we also know the PLA is interested in a long-term but potentially high-impact version of astroturfing. According to a 2018 article by Base 311 researchers that was likely intended to be a how-to guide for manipulating Facebook, the recommended approach is to integrate into preexisting online communities, participate in anodyne nonpolitical conversations and not draw too much attention, then (at the right time) inject the desired political narratives (Weng and Chen, 2018).[27] This tactic is exactly what we are worried generative AI will be especially good at.

For the Chinese military, generative AI offers the possibility to do something it could never do before: manipulate social media with human-quality content, at scale. Chinese military researchers routinely complain the PLA lacks the necessary amount of staff with adequate foreign-language skills and cross-cultural understanding.[28] Considering other examples of inauthentic content attributed to China, such as content produced during the 2019 Hong Kong protests, this deficit appears to be true of the Party-state broadly.[29] Open-source tools, such as Google Translate or Baidu Translate, can already solve this foreign-language-skill problem for Beijing if simple

direct translation was all that was needed. The real problem is that the PRC's Great Firewall has, in this narrow case, backfired to practically limit China's understanding of adversary (U.S.) society below the high-fidelity level the PLA needs to conduct the quality of social media manipulation that Russia did in 2016. Generative AI LLMs, such as ChatGPT, offer to bridge this cultural gap for the Party-state at scale. However, generative AI's reliance on massive amounts of training data will be a key focus for the PLA, and PLA information warfare researchers have even complained about the lack of internal data-sharing.[30]

PLA researchers already recognize the potential offered by generative AI, according to an initial review of *PLA Daily* articles through May 2023. Hu Xiaofeng, a top PLA researcher, paraphrased Friedrich Engels to say that "undoubtedly, the cutting-edge technology of AI represented by ChatGPT will inevitably be applied in the military field" (Hu, 2023). Specifically, Hu noted that for "cognitive domain operations, ChatGPT technology may also be used to produce fake news, fake emails, and even imitate human language styles for information deception, or be used in cyber attacks." This view was echoed in another article arguing that the "rapid development of generative AI and its wide application . . . is the general trend of cognitive warfare in the future" (Chen and Xu, 2023). Yet another said, "ChatGPT-like applications can efficiently generate massive amounts of fake news, fake pictures, and even fake videos to confuse the public . . . Compared with human beings, large-scale model technology has huge advantages in terms of quantity and time for its application toward information generation" (Shen, 2023). Another *PLA Daily* article argued that GPT-3 and emotional AI (情感智能 or 情感AI) are much better for social bots and public opinion guidance than deepfake technology, since

such AI "can learn a person's language style, and even act as that person to communicate with you. If you don't deliberately screen it, you can't judge its authenticity at all" (Wang, 2023).

At least some PLA researchers also understand the limitations. A 2023 *PLA Daily* article stated that the "full application of generative AI in the military field seems to be relatively far away," citing limited relevant training data, human trust issues for black-box models, and ethical challenges (Shen and Shu, 2023). Yet not all PLA views are positive: One 2023 *PLA Daily* article argued that the inevitable human bias introduced into ChatGPT by its U.S. creators presents a high risk that its outputs will have implicit bias toward Western political values, thereby subconsciously influencing (PRC) users (Chen and Xu, 2023).

One open question is whether the CCP and PLA will allow unorthodox narratives to be produced from generative AI, even if such narratives are exclusively for use abroad and likely to be successful at their intended purpose to influence foreign perceptions. CCP foreign propaganda is very often still constrained to domestically acceptable narratives and is sometimes crafted more for domestic audiences than foreign ones.[31] While PRC social media manipulation should, in theory, be able to escape this trap because of its often covert nature, the evidence suggests that PRC campaigns continue to center on CCP domestic propaganda narratives.[32]

PRC Intent Illustration: Li Bicheng

Li Bicheng, a Chinese military researcher who has likely helped the PLA operationalize AI for its information warfare and, specifically, social media manipulation, provides a useful illustration of what the PLA may be dreaming of with generative AI. Since at least 2016, Li has led a research effort to explore how to design

an operational system for "online public opinion struggle."[33] In a 2019 article as part of this effort, Li laid out a model for AI-enabled public opinion manipulation that matches the threat we have outlined above: a network of AI-controlled synthetic personae that are realistic enough to simulate public consensus on issues of concern to the CCP (Li, Hu, and Xiong, 2019). Li clearly revealed his intent in research published in 2016 that called for the PLA to improve its ability to conduct "online information deception" and "online public opinion guidance," the cornerstones of social media manipulation (Li, 2016).[34] Li's special importance within the PLA is evident in the fact that he coauthored his 2019 article with a researcher at Base 311, right after the unit was accused by Taiwan of election interference via social media (Beauchamp-Mustafaga and Drun, 2021).

While LLMs and text-to-image models were not available at the time of Li's research in 2019, Li accurately predicted the AI capability needed to overcome technical bottlenecks for maximally effective online public opinion warfare, writing that traditional social bots or trolls do not sound human enough to be fully effective. Specifically, Li complained that their "post generation is mechanized without regard for personality, occupation, and age differences; there is no individuality or simulation of human characteristics, so posts are

LI Bicheng, Chinese military researcher

PLA-affiliated researchers such as Li are proposing operating concepts for employing AI for social media manipulation.

easily identified and deleted," something generative AI elegantly solves (Li, Hu, and Xiong, 2019).

The model Li proposed has five main capabilities:

1. receive public opinion situation and guidance orders from command-and-control centers
2. select appropriate post generation models according to the topic, style, and emotional tone set by guidance orders, and generate posts with certain character traits
3. adjust guidance timing and methods based on current online public opinion
4. publish posts and conduct public opinion guidance based on set behavioral characteristics and guidance timing and methods
5. carry out coordinated online public opinion guidance between multiple intelligent agents (Li, Hu, and Xiong, 2019).

In the above five capabilities, generative AI will help with content generation (capability 2), and the potential for autonomous
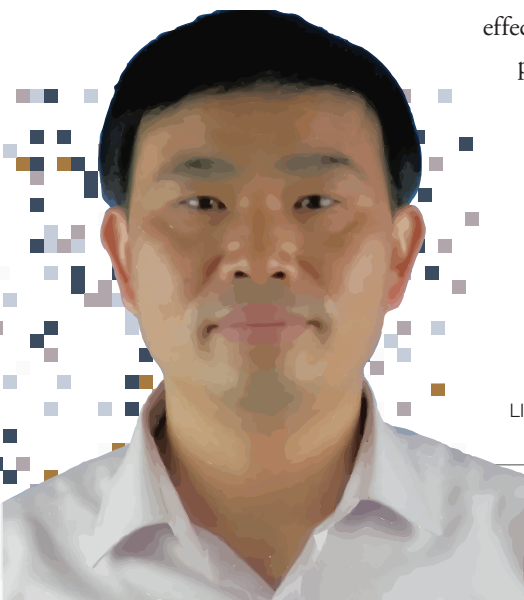
action from LLMs and their interactivity may mean they can assist with solving the orchestration (capability 3), delivery (capability 4), and coordination (capability 5) challenges. Li's 2020 article visualizes this model as a set of inputs (AI modeling and CCP objectives) that inform post (content) generation and post timing (delivery) mechanisms for outputs to social media, as shown in Figure 1.

Li's research is ongoing: He published another article in January 2023 on improving the outputs of a language model for better using emotion in text generation and thus generating more-convincing synthetic text (Li and Gong, 2023). Although we categorize this research as social media manipulation 2.0, since it is based on an earlier generation model (a fine-tuned version of Google's BERT), it clearly demonstrates Li's research interests evolving over time and seems ripe for leveraging generative AI for social media manipulation 3.0.

Thus, in addition to the PRC's growing technical capacity and well-documented intentions of using AI to spread propaganda, PLA-affiliated researchers such as Li are proposing operating concepts for employing AI for social media manipulation.

Illustrative Potential PRC Use Case: Taiwan

As the focal point of Chinese foreign social media manipulation, Taiwan illustrates the potential implications if China successfully adopts generative AI for its social media manipulation. Beijing is already known for using social media to interfere in Taiwanese politics since at least 2016. The CCP claims Taiwan and seeks to achieve unification through nonmilitary means if possible, and therefore it hopes to shape the outcome of Taiwanese elections toward Taiwanese politicians who Beijing views as being more favorable toward unification.[35]

FIGURE 1

PLA Researchers' Vision for Generative AI–Driven Social Media Manipulation
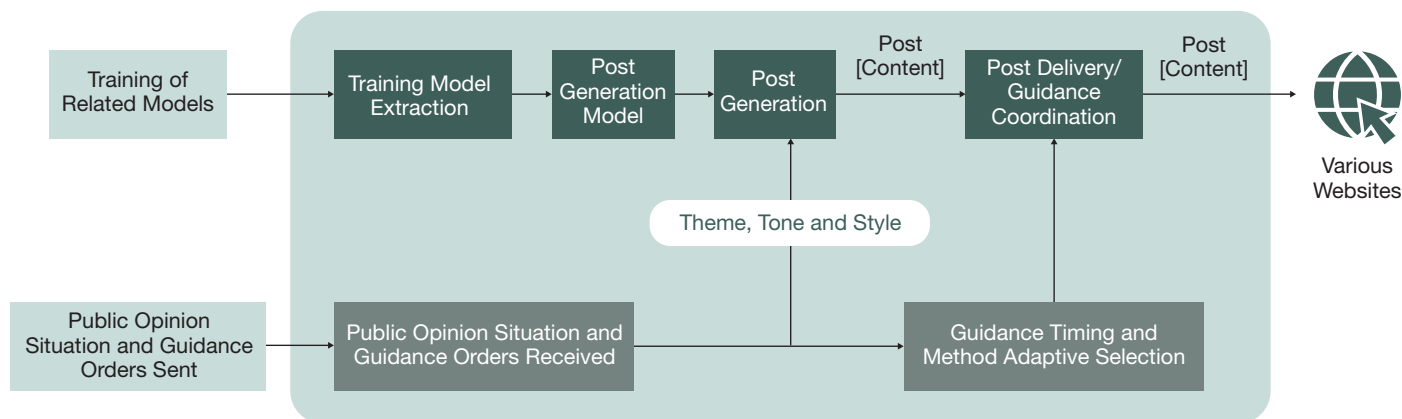
SOURCE: Adapted from Li et al., 2020.

China engages in widespread political interference against Taiwan, not just via social media manipulation but also by leaning on traditional media figures and cultivating election influencers via local power brokers, including criminal gangs, managed by the CCP's United Front Work Department (Cole, 2019; Hille, 2018). The first reports of Chinese election interference via social media manipulation against Taiwan emerged in 2018, when Taiwan's government accused the PLASSF of creating fake social media accounts and spreading fake news to interfere with the November 2018 elections, as practice for manipulating the 2020 elections in support of candidates with policies favorable to Beijing (Chung and Hetherington, 2018; Everington, 2018). The reported tactics were coordinated across media formats and largely centered on content generated by PRC traditional media; first, PRC state-owned media created videos and other content targeting Taiwan, then the PLASSF and other CCP-affiliated actors (e.g., the 50 Cent Army) spread the content on social media platforms (Chung and Hetherington, 2018; Everington, 2018).

PRC tactics for social media manipulation have evolved since 2018 to focus on more-organic content specifically tailored for social media, though broader PRC efforts still include influencing Taiwanese traditional media.[36] One shift involved a greater reliance on *content farms*: specifically, PRC-controlled websites that produce false or misleading content, which is spread by PRC-affiliated actors (and unwitting non-PRC users) on Taiwanese social media. Such tactics included cheapfakes and recontextualized media, and the PRC reportedly used bots to spread this content. Another shift involved PRC attempts to buy established Taiwanese social media accounts, whether Taiwanese media companies' official accounts or just popular accounts with large followings. Lastly, Beijing reportedly attempted to pay Taiwanese *influencers*, or online celebrities, to "advertise" pro-Beijing narratives within their normal content (Feng, 2019; Chen and Hetherington, 2023).

Despite these sophisticated, coordinated, and expensive efforts, foreign researchers have concluded that the PRC's attempts to influence Taiwan's 2020 election results were minimal at best. However, these attempts appear to have some measurable effects, worsening Taiwanese political and social polarization and widening perceived generational divides (Huang, 2020, p. 29). Social media manipulation 3.0 is likely to improve these Party-state efforts meaningfully.

Looking forward, the theoretical benefits of generative AI may be borne out by PRC efforts against Taiwan. The prospect

With generative AI, PRC malign actors will likely be able to appear more authentic, simply using few-shot prompt generative AI to produce text content for social media posts.

of improved authenticity will be the most important, since previous CCP attempts have been generally regarded in Taiwan as ultimately not very effective because they are not very tailored and thus (so far) have been easy to spot, such as using the wrong Chinese-language script (simplified Chinese, used on the mainland, instead of traditional Chinese, used in Taiwan; Harold, Beauchamp-Mustafaga, and Hornung, 2021). With generative AI, PRC malign actors will likely be able to appear more authentic, simply using few-shot prompt generative AI to produce text content for social media posts. Indeed, the director-general of Taiwan's main intelligence organ, the National Security Bureau, warned Taiwanese lawmakers in April 2023, "We are closely watching whether [China] will use new generative AI applications in disseminating disinformation" (Shan, 2023). Going a step further, generative AI that is fine-tuned on Taiwanese social media—easily attainable by the PRC government—could revolutionize this content. Such AI would improve overt propaganda, such as China Global Television (a key PRC overseas propaganda organization) and, more importantly, would make astroturfing more convincing. It is unclear exactly how much of a constraint human labor was on PRC efforts previously; regardless, generative AI will directly benefit Beijing by reducing its labor requirements. This, in turn, will be helpful for expanding the scale and reach of PRC efforts, especially for astroturfing. Lastly, improved authenticity will likely decrease the possibility that PRC efforts are detected by Taiwan, further improving the impact of PRC efforts by going unnoticed. Building on the general tactics and implications outlined in Table 2, Table 3 illustrates how generative AI may change PRC social media manipulation tactics against Taiwan in the future.

TABLE 3

How Does Generative AI Change the Game for Chinese Social Media Manipulation Against Taiwan?

Common PRC Tactic	Definition	Previous Shortcoming	Potential Implication with Generative AI
Advertising	Paid promotional content to support a cause or actor	PRC paid Taiwan influencers to promote pro-CCP content, but they sometimes were easy to identify with blatant one-off messages	May diminish; PRC no longer needs to pay others to create viral content if it is able to generate convincing, authentic text
Bots for astroturfing	Using large numbers of inauthentic (fake) accounts (bots) to create the appearance of a broad consensus on a topic	PRC has largely relied on human-generated comments, limiting quality and scale	Likely to increase dramatically; generative AI will give bots written voices that are near-indistinguishable from human-created content
Cheapfakes and recontextualized media	Supporting a campaign either with simple edits or by repurposing media (usually, images)	PRC attempts are relatively easy to identify and slow enough for Taiwan government to expose and debunk	May diminish; realistic, highly believable fakes will be far cheaper to make en masse and may not be able to be identified or may overwhelm Taiwan government response capabilities
Impersonation	Pretending to be another person in order to misrepresent their position or views	PRC relies on pressuring public individuals into creating misleading information (especially, confessions)	Now possible to (1) mass-generate text in the style of a given individual's writing (2) falsify images of an individual and produce those images en masse. There is no longer a need to actually coerce a targeted individual
Keyword squatting	Creating mass content to manipulate search engine results related to a given term, phrase, or hashtag	Past PRC campaigns on Xinjiang issues have lacked variety, making them easier to detect	Generative AI does not revolutionize keyword squatting's mechanism but permits squatters to automate mass content generation containing a given keyword
Swarming	Loosely organized groups coordinating to fill an information space (e.g., spamming a comment section)	50 Cent Army members are inconsistent in their ability to avoid detection or achieve specific narrative goals	Generative AI automates the process of creating mass unique content for spamming a comment section or otherwise drowning out a narrative
Testimonials	Personal stories used to elicit emotional reactions or sway opinions	PRC-manufactured testimonials have historically been presented on state media and appear to be relatively scripted	Generative AI is capable of writing short-form and long-form testimonials of wide-ranging content on a mass scale, representing various demographics for both broad and niche effects

SOURCES: Adapted from Harvard Kennedy School Shorenstein Center for Media, Politics, and Public Policy, 2022; and Huang, 2020.

What Are the Limitations?

Although we believe that social media manipulation using generative AI poses a meaningful threat, we also acknowledge there are limitations to generative AI. Despite the "AI" moniker, LLMs and text-to-image models do not appear to have humanlike, general intelligence. They are very data-rich representations of patterns in language and images, but fundamentally there is no understanding involved. Output from LLMs such as ChatGPT is often inaccurate or contextually inappropriate enough that attentive readers can detect something is off, because, at their core, LLMs are statistical models of next-word prediction. Similarly, text-to-image outputs often have visible anomalies (e.g., hands with seven fingers) because these models are only statistical impressions of things humans understand holistically and accurately (such as people). Generative AI is not perfect, nor is it undetectable by humans (yet). We do note, however, that the quality of these models is rapidly improving.

Looking ahead to more-sophisticated AI, we note that there are technical limitations to building improved, next-generation models and to improving some of their basic architectural limitations. Foundational LLMs such as GPT-3 and GPT-4 are not cheap to build; the latter reportedly cost $100 million to develop (Knight, 2023). Deploying them can also be extremely costly: While we do not have exact figures, running ChatGPT might cost $100,000 per day, given the cost of renting high-end graphics cards via cloud computing services.[37] Although we cannot predict pricing for computing power in the future, training the next generation of larger and higher-performance models is still likely to be very expensive: Fine-tuning existing models is cheap, but building new ones is a nontrivial endeavor. This cost may be prohibitive to some nonstate actors; however, if the Chinese mili-

Running ChatGPT might cost $100,000 per day, given the cost of renting high-end graphics cards via cloud computing services.

tary leadership decides to pursue such models, China can certainly afford it, with its military budget estimated at more than $290 billion in 2021.[38] We do note, however, that the technical limitations on cost for building and deployment are specific to LLMs: Text-to-image models, such as Stable Diffusion, are much cheaper to train and can be deployed on personal computers.[39]

Improving the performance of LLMs is also limited by architecture, such as the length of text inputs during modeling. Although humans are relatively slow and forgetful compared to machines, we have great scope in how we make sense of text data. A reader may not remember every word of *The Lord of the Rings*, but they can make sense of the whole story (for example, understanding character arcs and relationships that span the entire narrative). However, the current generation of LLMs uses transformer architectures that can process a limited amount of text at a time, though even this limitation is changing rapidly: When we started drafting this report in February 2023, LLMs could only ingest two to four pages of text; several months later,

that has grown to eight pages and even, reportedly, 40 pages (Hern and Bhuiyan, 2023). Transformers are quadratic in how they ingest text: Doubling the length of the input quadruples the computing requirements, and this can quickly scale out of feasibility (Dubey, 2021). There is ongoing research in this area, but LLMs as of May 2023 can only make inferences on a model built from (relatively) short chunks of documents, which appears to limit their accuracy and appropriateness for various tasks. Indeed, Microsoft blamed the early failures of its Bing version of GPT-4 on this problem (Weise, 2023).

Finally, as much as generative AI appears to be revolutionary for social media manipulation, it is not clear just how much real-world impact social media influence campaigns have in the first place, regardless of how convincing they might be. While there is a body of evidence showing some effects from such campaigns, the degree and duration is not clear. Some research has found that social media manipulation campaigns do not change strongly held opinions (Cohen et al., 2021), and other research has shown how little influence some campaigns have (e.g., pro-Western campaigns attributed to the U.S. government; Graphika and Stanford Internet Observatory, 2022). There is reason to think that malign information operations, including social media manipulation, have meaningful effects: Certainly, U.S. adversaries conducting these operations think so. But we do not have clarity on how serious the threat is.

## What Should Be Done About Generative AI?

Although generative AI presents threats in terms of social media manipulation, a coherent, proactive response by the U.S.

government and broader technology and policy community may mitigate those threats. There are technical and policy solutions to specific aspects of the larger problem. Additionally, the U.S. national security establishment has a role to play. Finally, diplomacy and engagement with both adversaries and allies may be critical to mitigating the informational threat of generative AI. Please note we do not discuss mitigation strategies that are country- or platform-specific. Regulating U.S. or European Union AI technology or placing guardrails in high-profile models, such as ChatGPT, will not have any effect on how Russia might use the open-source YaLM 100B Russian-language LLM or how the PLA might leverage PanGu-Alpha. *Mitigation strategies must account for generative AI being ubiquitous and unregulated globally.*

Many of the mitigation strategies we raise here involve trade-offs, whether between profits and platform integrity for private companies or between freedom of speech and information security for the U.S. government and broader public. These public and private policy decisions may be fraught, but they will be improved by an inclusive and informed public conversation that begins now, not after another foreign (or domestic) attack on the U.S. democratic process in the 2024 elections. We do not recommend any specific solutions as easy choices but instead raise these options for consideration to start the conversation.

Technical Mitigations

The U.S. government, social media platforms, and the AI industry can explore technical solutions to generative AI threats, but foreseeable mitigations in the near term all appear to fall short right now. The overall principle should be to identify, attribute, and remove inauthentic personae (accounts) in order to restore
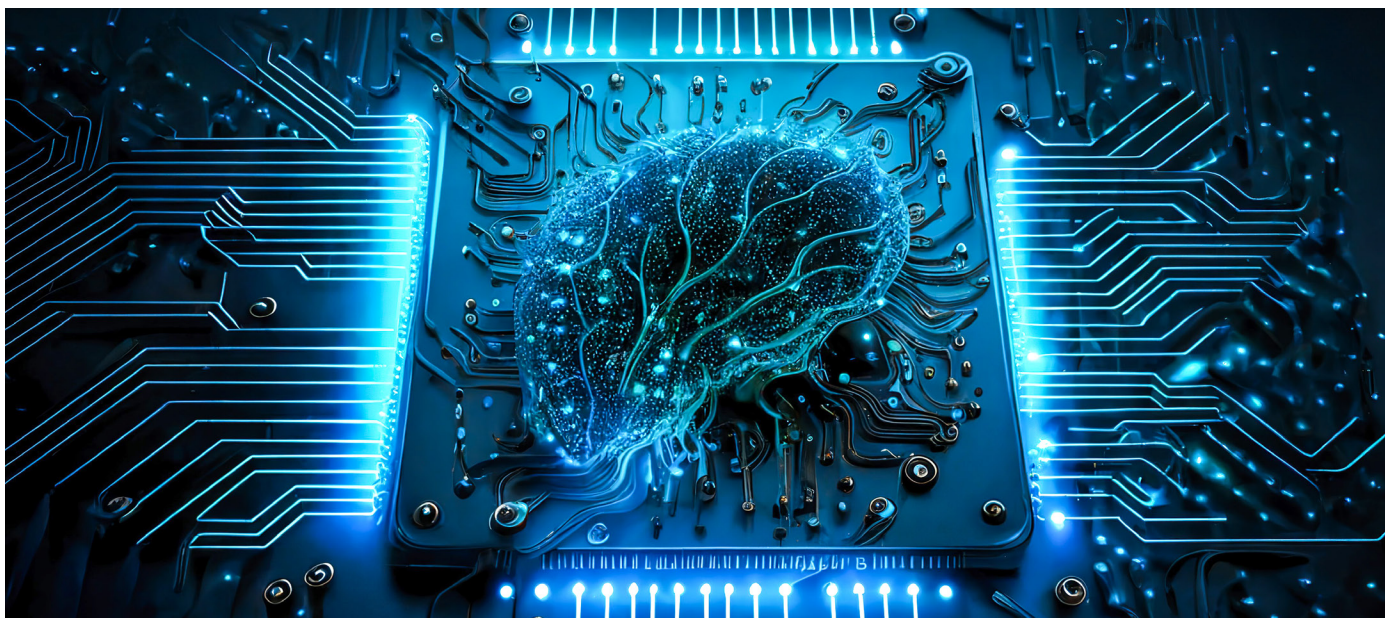
faith and trust in the public conversation on social media. While this may be a moment in which the offense-defense balance shifts toward the offense, offensive advantage has historically been transitory; thus, policymakers should not despair that the advent of generative AI means social media manipulation is forever impossible to counter.[40] The United States can invest in defensive technology that can detect inauthentic content at scale or in watermarking technology to verify the source of the content. At this time, it is not clear what technologies may be available and how effective they will be. In response to Chat-GPT's potential use by students, some vendors are touting "AI detection" tools, but these are generally very crude models that use simple perplexity measures (essentially, word variability) and are trivial to defeat. Even OpenAI's tool to detect its own model

is fooled simply by asking ChatGPT to do so (e.g., "use a variety of words with a perplexity of about 600"). That does not mean, however, that better approaches and thus better detection models cannot be developed.

If the current prospects for detecting generative AI–produced content are slim, then it is also worth considering how to address another major part of the problem: content delivery. With the improvements offered by generative AI for content creation, content delivery (specifically, inauthentic accounts) may be the new limiting factor for social media manipulation. Platforms should redouble their efforts to make it harder for new accounts to be opened by malign actors and should redouble efforts to identify and remove inauthentic accounts.

Short of directly identifying and removing inauthentic content, another option would be to bolster public confidence in authentic social media content. A blockchain system for media that acts as a kind of public chain of custody and authenticity may help create confidence that an image or document is authentic and comes from when and where it claims (Horvitz, 2022). Such a system would make alterations to documents or images visible and could highlight inauthentic documents as having no verifiable, trustworthy lineage. Similarly, journalistic reputational systems via blockchain may give the public a way to quickly verify trustworthy sources (Almasoud, Hussain, and Hussain, 2020).

Other technical measures to restrict the proliferation of generative AI–produced content could include attempting to restrict adoption by malign actors. While AI models will be hard to contain or control, the underlying hardware used to train current and future language-generation models may be easier to control.[41] These models are nontrivial to produce: They require massive datasets, computing power, and skilled technical labor to train them (Goldstein, Sastry, et al., 2023), so the sort of restrictions on access to AI hardware the U.S. government enacted against China and Russia in 2022 may help. We note, however, that prohibitions may be hard to enforce, and the United States likely cannot ban every potential bad actor—Iran, for example—from developing and exploiting this new generation of language AI.

Policy Mitigations

While we do not think that U.S. or Western regulation of generative AI can restrict foreign social media manipulation efforts, we do think that there are regulatory steps the United States can take to make it harder for adversaries to gain access to platforms and remain hidden. Regulating social media platforms to create norms of transparency, public accountability, and access for researchers might be a powerful tool against social media manipulation. We think regulatory policy that lays out threats and platform responsibilities, while protecting civil liberties, may be an important first step toward reducing platforms' vulnerability as vectors of transmission for social media manipulation (Rochefort, 2020; West, 2017). Further, regulation that requires transparency and access for researchers could be a critical way to strengthen democracies against these threats (Aral and Eckles, 2019).

Another potentially powerful (but complicated) policy intervention would be requiring platforms to verify the identity of users behind accounts (Balasubramaniam, 2020; De Leon, Enriquez, and Tiglao, 2019), creating a uniform standard for social media platforms. On the one hand, requiring identity verification, similar to verification required for banking, could directly attack the scale advantage of generative AI social media manipulation. On the other hand, identity verification may have a chilling effect on free speech: The anonymity of social media helps support unpopular speech (for example, government criticism). We note that this is primarily a domestic mitigation strategy for U.S.-based platforms. Autocratic regimes such as Russia or China would still, of course, be free to flood and manipulate platforms within their influence, such as WeChat, TikTok, and VKontakte.

In addition to internally facing regulation for platforms, outward-facing national security policy can help mitigate the threat of social media manipulation. The U.S. Department of Defense (DoD) must be prepared to operate across a con-

tested information environment in which generative AI scales propaganda and makes it ubiquitous, preparing across doctrine, organization, training, materiel, leadership and education, personnel, facilities, and policy. An example is force-wide training and education, for which DoD as a whole lacks sufficient resources for operations in contested information environments: specifically, vignette or case study repositories, access to subject-matter experts, simulated training environments, and supporting infrastructure (U.S. Government Accountability Office, 2023). Preparations to deal with generative AI could include multiple DoD components, including the Under Secretary of Defense for Personnel and Readiness, Under Secretary of Defense for Policy, Chairman of the Joint Chiefs of Staff, armed services, and combatant commands. Likewise, DoD components could robustly research and red-team possible adversary use of generative AI, informing national security policy. One example is educating the joint force on identifying images produced by generative AI: Although such images may be photorealistic, they often have visual anomalies (e.g., too many fingers) that can be detected with a close eye.

Lastly, if the detection of generative AI content will be much more challenging, then public attribution will also be difficult even as it becomes more important. The policy community should discuss standards for attribution now, and the United States and other interested governments should consider how they might support public attribution of malign actors. Another step is to raise public awareness of this threat and to expand ongoing public education and media literacy efforts to include addressing generative AI (Huguet et al., 2021).

Diplomatic Mitigations

The United States could also consider engaging in dialogue with China on the topic of generative AI and the risks for social media manipulation. Although any dialogue is challenging, given ongoing tensions in the bilateral relationship, Track II dialogues may help facilitate at least some conversation between the two sides' broader policy and research communities. The objective of these conversations would be to better understand how each side views the risks of generative AI and whether there is room for agreement on restraining government use and limiting malign uses by domestic nonstate actors.

Short of direct or even indirect engagement with China, Russia, and others on rules of the road for nation-state uses of generative AI, the U.S. government could also (at a minimum) make a public declaration of its principles for using this emerging capability. Such a declaration would be similar to the February 2023 declaration by the U.S. Department of State on the "Responsible Military Use of Artificial Intelligence and Autonomy" (U.S. Department of State, 2023). China made a similar declaration in December 2021 (Ministry of Foreign Affairs of the People's Republic of China, 2021b). While it would serve the immediate purpose of clarifying U.S. policy, it would also ideally serve to allay some concerns in Beijing, Moscow, and elsewhere, though this is unlikely in practice due to a lack of mutual trust with the Chinese and Russian governments. It might also inspire other countries to consider adopting similar principles and might at least limit the proliferation of nation-state and domestic non-state actors employing generative AI for malign purposes.

The United States should also begin monitoring for PRC employment of generative AI–produced content as an indicator of intent and technological progress. While early PRC employ-

ment may be experimental and poor, it would confirm PRC intent to leverage this emerging capability and provide an opportunity for the United States and others to raise awareness. Special focus should be paid to Chinese efforts against Taiwan, which has historically been Beijing's testing ground for information warfare, including social media manipulation. Such efforts would benefit from engagement with Taiwan on the topic, detailed below. One specific indicator of Chinese intent would be PRC development of LLMs that focus on Taiwanese text, such as pulling from PTT (Taiwan's version of Reddit).

Multilaterally, the United States should begin engaging with allies and partners (especially, Taiwan) on these emerging risks, evidence of Chinese employment, and potential countermeasures. While PRC social media manipulation has historically been a limited concern outside Taiwan and the United States, generative AI has the potential to extend China's capability to a much wider range of target countries, such as Japan, South Korea, and the Philippines, as well as other countries in Southeast Asia and Europe. Although raising awareness and sharing information—especially on evidence of Chinese employment of generative AI—would be the first goal, reaching consensus on norms of behavior and cooperation on countermeasures would be good medium- and long-term goals. Such conversations can be folded into common 2+2 bilateral diplomacy and defense dialogues or multilateral forums, such as the Global Cooperation and Training Framework, which provides a platform for Taipei to showcase its expertise on the topic.

Lastly, the United States could consider the potential of arms control for social media manipulation–related capabilities (especially, generative AI, as described in this Perspective). While arms control in general is facing headwinds amid severe downturns in relations with China and Russia, especially given the fraught recent history of nuclear arms control and failed attempts at cyber arms control, the prospects of a highly destructive offensive-favoring but unregulated new capability (generative AI) may make the topic worthy of consideration. In fact, China and Russia have historically favored, at least in public statements, arms control for what they call "information weapons."[42] If nothing else, some of the low-hanging fruit from the arms control process, such as confidence-building measures, may be worthy of further research and policy consideration.

Conclusion: A Proactive, Broad Strategy

We are at the start of a new era of potential social media manipulation. Many of the former constraints on malign influence activities over social media (particularly, the trade-off between scale and quality) appear to be largely or even completely obviated by advances in generative AI. Further, these advances in AI are continuing at an explosive pace, not only in terms of new and improving generative capabilities but also in terms of emerging capabilities for AI-enabled distribution and management. The U.S. government and broader technology and policy community should respond proactively, considering a variety of mitigations to lessen potential harm. Although we have emphasized and unpacked here the specific intent and interests of China vis-à-vis Taiwan, such concerns extend to a variety of malign state and nonstate actors. Therefore, we strongly suggest the development of a coherent, proactive, and broad strategy for dealing with this new threat.

Endnotes

[1] We define *social media manipulation* as the artificial intervention to influence discourse on social media platforms, whether by state or nonstate, domestic or foreign actors. Strictly speaking, the technology we address here is machine learning, a subset of AI. However, popular discourse uses the two terms somewhat interchangeably, so we default to *AI* here.

[2] Early text-to-video models have been released at the time of this writing but are still not as mature or ready for deployment as text and image generative models.

[3] Russia's Internet Research Agency is an online troll farm that engages in influence campaigns and election interference on behalf of the Russian government. The name for China's 50 Cent Army originates from the notion that online commentators were paid RMB¥0.50 per post to spread pro-CCP propaganda.

[4] Estimates from the mid-2010s suggest that the typical American has 15 to 20 online friends who they have never met in person (Center for the Digital Future, 2019).

[5] For example, see Goldstein, Sastry, et al., 2023; and International Institute for Strategic Studies, 2023.

[6] For an overview of the disinformation process, see Sedova et al., 2021a.

[7] Although deepfakes have been produced with generative adversarial networks and thus technically are generative AI content, we argue that the newer generation of LLMs and other models are substantially different.

[8] For a good overview of the technology underlying deepfakes, see Hwang, 2020.

[9] For earlier work on this topic, see Sedova et al., 2021b; Helmus, 2022; Goldstein, Sastry, et al., 2023; and OpenAI, 2022.

[10] See, for example, the AI-generated "photo" of a couple on a carousel in Yang, 2023.

[11] ChatGPT has a limited "context memory" (approximately 3,000 words) in any given conversation, so few-shot learning while interacting with the model is temporary, forgotten after about 3,000 words have passed. In contrast, these models can be fine-tuned with additional training and data that permanently changes the model and its output. For example, OpenAI has already made GPT-3 available for fine-tuning (see OpenAI, undated) and may do so for ChatGPT soon as well.

[12] See, for example, Edwards, 2022; and Liu et al., 2022.

[13] See, for example, a deepfake of President Joseph Biden attacking transgender people (Lajka, 2023).

[14] For example, see François, Nimmo, and Eib, 2019.

[15] For a good, recent review of PRC-attributed social media manipulation, see Zhang, Hoja, and Latimore, 2023.

[16] Russia has also made use of early generative AI for profile photos (see Grossman et al., 2021).

[17] A May 2023 report by a Chinese government–affiliated think tank claimed that various Chinese organizations had developed 79 LLMs since 2020. See Li and Baptista, 2023.

[18] For a recent comparison between Chinese and foreign LLMs, see Ding, 2023.

[19] A search of state media Xinhua and the Chinese Communist Party's official newspaper *People's Daily* for "ChatGPT" from November 30, 2022 (when ChatGPT was released), to January 13, 2023, returned only four unique reports.

[20] For background on CCP concerns, see Delury, 2022.

[21] For more on PRC threat perceptions from social media and the Arab Spring, see Beauchamp-Mustafaga and Chase, 2019; and Zhao, 2018.

[22] For some PLA views, see Chen, 2016; and Zeng and Shi, 2014, pp. 79–81. For a PRC Ministry of Foreign Affairs view, see Ministry of Foreign Affairs of the People's Republic of China, 2023.

[23] For earlier research on PLA interest in emerging technologies for information warfare, see Chen, 2022; and Chen, 2023. For a good, recent review of multiple PRC actors involved in PRC influence operations, see Zhang, Hoja, and Latimore, 2023.

[24] For more on the PLA's assigned missions, see State Council Information Office of the People's Republic of China, 2019.

[25] For more on the Three Warfares and public opinion warfare, see Wu and Liu, 2014.

[26] Early PLA sources with this intent include Ji, Sun, and Yu, 2005; and Yang and He, 2007.

[27] The authors list their affiliation as Huayi Broadcasting Corporation (中国华艺广播公司), but this is a well-known front organization for Base 311. See Beauchamp-Mustafaga and Drun, 2021.

[28] The most prolific PLA information warfare researcher lamenting the PLA's lack of foreign language skills is Liang Xiaobo. See, for example, Li and Liang, 2018, pp. 1–6; and Liang, 2019.

[29] For more on the shortcomings of 2019 Hong Kong–related disinformation, see Conger, 2019; and Dotson, 2019.

[30] See, for example, Liu and Zhang, 2016.

[31] See, for example, Cowhig, 2021.

[32] For foreign propaganda, see Ryan et al., 2021. For one PRC inauthentic social media campaign, see Strick, 2021.

[33] Notable PRC government funding includes National Social Science Fund grants 14BXW028 and 19BXW110. For key articles in this general line of research, see Li, 2016; Li, Hu, and Xiong, 2019; and Li et al., 2020.

[34] Li lists his affiliation as Huaqiao University (华侨大学), but he is a career PLA researcher, and we assess that he likely still maintains his PLA ties, based on his coauthored research with PLA organizations.

[35] For some research on this practice, see Beauchamp-Mustafaga and Drun, 2021; Harold, Beauchamp-Mustafaga, and Hornung, 2021; and Insikt Group, 2020.

[36] For two reviews of PRC tactics during Taiwan's 2020 elections, see Insikt Group, 2020; and Huang, 2020. For reporting on broader PRC efforts targeting Taiwanese media, see Lee and Cheng, 2019.

[37] See, for example, Goldstein, 2022.

[38] Data from the Stockholm International Peace Research Institute via Center for Strategic and International Studies, 2023.

[39] The CEO of Stability AI reported the cost of training Stable Diffusion at $600,000 (Mostaque, 2022).

[40] For more on offense-defense balance, see Brown et al., 2004.

[41] Fine-tuning and deploying image generation models is comparatively easy, and AI hardware bans are unlikely to have any effect on their proliferation and deployment.

[42] See, for example, Ministry of Foreign Affairs of the People's Republic of China, 2011; and Ministry of Foreign Affairs of the People's Republic of China, 2017. For relevant research, see Farnsworth, 2011; Barmin et al., 2011; and McKune, 2015.

References

Almasoud, Ahmed S., Farookh Khadeer Hussain, and Omar K. Hussain, "Smart Contracts for Blockchain-Based Reputation Systems: A Systematic Literature Review," *Journal of Network and Computer Applications*, Vol. 170, November 15, 2020.

Aral, Sinan, and Dean Eckles, "Protecting Elections from Social Media Manipulation," *Science*, Vol. 365, No. 6456, August 30, 2019.

Balasubramaniam, Nandhagopal, "Blockchain Based Digital Identity Verification for Social Media," *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences*, Vol. 7, No. 8, 2020.

Barmin, Yury, Grace Jones, Sonya Moiseeva, and Zev Winkelman, "International Arms Control and Law Enforcement in the Information Revolution: An Examination of Cyber Warfare and Information Security," *Connections*, Vol. 10, No. 4, Fall 2011.

Beauchamp-Mustafaga, Nathan, "Cognitive Domain Operations: The PLA's New Holistic Concept for Influence Operations," *China Brief*, Vol. 19, No. 16, September 6, 2019.

Beauchamp-Mustafaga, Nathan, *Chinese Next-Generation Psychological Warfare: The Military Applications of Emerging Technologies and Implications for the United States*, RAND Corporation, RR-A853-1, 2023. As of June 20, 2023:
https://www.rand.org/pubs/research_reports/RRA853-1.html

Beauchamp-Mustafaga, Nathan, and Michael S. Chase, *Borrowing a Boat Out to Sea: The Chinese Military's Use of Social Media for Influence Operations*, Johns Hopkins School of Advanced International Studies Foreign Policy Institute, 2019.

Beauchamp-Mustafaga, Nathan, and Jessica Drun, "Exploring Chinese Military Thinking on Social Media Manipulation Against Taiwan," *China Brief*, Vol. 21, No. 7, April 12, 2021.

Brown, Michael E., Owen R. Coté, Jr., Sean M. Lynn-Jones, and Steven E. Miller, eds., *Offense, Defense, and War*, MIT Press, 2004.

Bu Jiang [卜江], Lao Songyang [老松杨], Bai Liang [白亮], Guo Xiaoyi [郭小一], and Liu Haitao [刘海涛], "The Research on Video Based Psychological Warfare and Its Key Technology" ["基于视频的心理战及其关键技术"], *Fire Control and Command Control* [火力与指挥控制], Vol. 36, No. 12, December 2011.

Center for Strategic and International Studies, "What Does China Really Spend on Its Military?" updated May 8, 2023.

Center for the Digital Future, "Web Insight: Do You Have Online Friends You Have Never Met in Person?" October 7, 2019.

Charon, Paul, and Jean-Baptiste Jeangène Vilmer, *Chinese Influence Operations: A Machiavellian Moment*, Institute for Strategic Research of the French Ministry for the Armed Forces, October 2021.

Chen Dongheng [陈东恒] and Xu Yan [许炎], "Generative AI: A New Weapon for Cognitive Confrontation" ["生成式AI: 认知对抗的新武器"], *PLA Daily*, April 4, 2023. As of June 16, 2023:
http://www.81.cn/szb_223187/szbxq/index.html?paperName=jfjb&paperDate=2023-04-04&paperNumber=07&articleid=902532

Chen, John, "China's Cyber Capabilities: Warfare, Espionage, and Implications for the United States," testimony presented before the U.S.-China Economic and Security Review Commission, Exovera Center for Intelligence Research and Analysis, February 17, 2022.

Chen, John, "Cyber and Influence Operations," in William C. Hannas and Huey-Meei Chang, eds., *Chinese Power and Artificial Intelligence: Perspectives and Challenges*, Routledge, 2023.

Chen Qingbao [陈庆宝], "Analysis on the U.S. 'Astroturfing' and Research on Countermeasures" ["对美国' 网络水军' 的分析及对策研究"], *Military Correspondent* [军事记者], No. 11, 2016.

Chen Yu-fu and William Hetherington, "Influencers' Funding Needs Scrutiny: Researcher," *Taipei Times*, January 28, 2023.

Cheng, Evelyn, "China's A.I. Chatbots Haven't Yet Reached the Public Like ChatGPT Did," *CNBC*, April 28, 2023.

"China Mandates Security Reviews for AI Services Like ChatGPT," *Bloomberg*, April 11, 2023.

Chung Li-hua and William Hetherington, "China Targets Polls with Fake Accounts," *Taipei Times*, November 5, 2018.

Cohen, Raphael S., Nathan Beauchamp-Mustafaga, Joe Cheravitch, Alyssa Demus, Scott W. Harold, Jeffrey W. Hornung, Jenny Jun, Michael Schwille, Elina Treyger, and Nathan Vest, *Combating Foreign Disinformation on Social Media: Study Overview and Conclusions*, RAND Corporation, RR-4373/1-AF, 2021. As of February 1, 2023:
https://www.rand.org/pubs/research_reports/RR4373z1.html

Cole, J. Michael, "More Than 70 Participants from Taiwanese Media Attend 4th Cross-Strait Media Summit in Beijing," *Taiwan Sentinel*, May 11, 2019.

Conger, Kate, "Facebook and Twitter Say China Is Spreading Disinformation in Hong Kong," *New York Times*, August 19, 2019.

Cox, Joseph, "Facebook's Powerful Large Language Model Leaks Online," *Motherboard*, March 7, 2023.

Cowhig, David, "2021: Propagandizing Foreigners to Reach China's Domestic Audience: Why PRC External Propaganda Is Often Ineffective," *David Cowhig's Translation Blog*, November 7, 2021. As of March 1, 2023: https://gaodawei.wordpress.com/2021/11/07/propgandizing-foreigners-to-reach-chinas-domestic-audience-why-prc-external-propaganda-is-often-ineffective/

De Leon, Jomari James T., Keir Cedric L. Enriquez, and Jose Angelo C. Tiglao, "Rise of the Troll: Exploring the Constitutional Challenges to Social Media and Fake News Regulation in the Philippines," *Ateneo Law Journal*, Vol. 64, No. 1, August 2019.

Delury, John, *Agents of Subversion: The Fate of John T. Downey and the CIA's Covert War in China*, Cornell University Press, 2022.

Ding, Jeffrey, "ChinAI #231: Latest SuperCLUE Rankings of Large Language Models," *ChinAI*, July 31, 2023. As of August 1, 2023: https://chinai.substack.com/p/chinai-231-latest-superclue-rankings

Dotson, John, "Chinese Covert Social Media Propaganda and Disinformation Related to Hong Kong," *China Brief*, Vol. 19, No. 16, September 6, 2019.

Dubey, Avinava, "Constructing Transformers for Longer Sequences with Sparse Attention Methods," *Google Research* blog, March 25, 2021. As of February 3, 2023: https://ai.googleblog.com/2021/03/constructing-transformers-for-longer.html

Edwards, Benj, "AI Image Generation Tech Can Now Create Life-Wrecking Deepfakes with Ease," *Ars Technica*, December 9, 2022.

Everington, Keoni, "China's 'Troll Factory' Targeting Taiwan with Disinformation Prior to Election," *Taiwan News*, November 5, 2018.

Farnsworth, Timothy, "China and Russia Submit Cyber Proposal," *Arms Control Today*, November 2011.

Feng, Coco, "Chinese Tech Firms Take Heed of Country's Strict Online Moderation as They Rush to Bring Their ChatGPT-Like Services to Market," *South China Morning Post*, February 11, 2023.

Feng, Emily, "Taiwan Gets Tough on Disinformation Suspected from China Ahead of Elections," *NPR*, December 6, 2019.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini, "The Rise of Social Bots," *Communications of the ACM*, Vol. 59, No. 7, July 2016.

Fielding, Nick, and Ian Cobain, "Revealed: US Spy Operation That Manipulates Social Media," *The Guardian*, March 17, 2011.

François, Camille, Ben Nimmo, and C. Shawn Eib, *The IRACopyPasta Campaign*, Graphika, October 2019.

Garnaut, John, "Engineers of the Soul: Ideology in Xi Jinping's China by John Garnaut," Sinocism, January 16, 2019.

Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz, "Can AI Write Persuasive Propaganda?" *SocArXiv*, April 8, 2023.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," *arXiv*, January 10, 2023.

Goldstein, Tom [@tomgoldsteincs], "How many GPUs does it take to run ChatGPT? And how expensive is it for OpenAI? Let's find out!" Twitter post, December 6, 2022. As of June 16, 2023: https://twitter.com/tomgoldsteincs/status/1600196981955100694

Graphika, *Deepfake It Till You Make It: Pro-Chinese Actors Promote AI-Generated Video Footage of Fictitious People in Online Influence Operation*, February 2023.

Graphika and Stanford Internet Observatory, *Unheard Voice: Evaluating Five Years of Pro-Western Covert Influence Operations*, Stanford Digital Repository, August 24, 2022.

Grossman, Shelby, Renée DiResta, Khadeja Ramali, Rajeev Sharma, Samantha Bradshaw, and Karen Nershi, *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.

Gui Qing, Koh, and John Shiffman, "Beijing's Covert Radio Network Airs China-Friendly News Across Washington, and the World," Reuters, November 2, 2015.

Guo Quanzhong [郭全中], "How to Build Smart Media—Taking the 'Artificial Intelligence Editorial Department' of China Media Group as an Example" ["智媒体如何打造—以中央广播电视总台 '人工智能编辑部' 为例"], *Youth Journalist* [年轻记者], February 2020.

Harold, Scott W., Nathan Beauchamp-Mustafaga, and Jeffrey W. Hornung, *Chinese Disinformation Efforts on Social Media*, RAND Corporation, RR-4373/3-AF, 2021. As of February 22, 2023: https://www.rand.org/pubs/research_reports/RR4373z3.html

Harvard Kennedy School Shorenstein Center for Media, Politics, and Public Policy, *The Media Manipulation Casebook Code Book*, version 1.4, updated January 7, 2022.

Hee Song, Chan, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su, "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models," *arXiv*, December 8, 2022.

Helmus, Todd C., *Artificial Intelligence, Deepfakes, and Disinformation: A Primer,* RAND Corporation, PE-A1043-1, July 2022. As of June 16, 2023: https://www.rand.org/pubs/perspectives/PEA1043-1.html

Hendrix, Justin, and Dan Morozoff, "Media Forensics in the Age of Disinformation," in Husrev Taha Sencar, Luisa Verdoliva, and Nasir Memon, eds., *Multimedia Forensics*, Springer, 2022.

Hern, Alex, and Johana Bhuiyan, "OpenAI Says New Model GPT-4 Is More Creative and Less Likely to Invent Facts," *The Guardian*, March 14, 2023.

Hille, Kathrin, "China's 'Sharp Power' Play in Taiwan," *Financial Times*, November 20, 2018.

Horvitz, Eric, "On the Horizon: Interactive and Compositional Deepfakes," *Proceedings of the 2022 International Conference on Multimodal Interaction*, Association for Computing Machinery, November 2022.

Hu Xiaofeng [胡晓峰], "How Should We View ChatGPT?" ["ChatGPT我们该怎么看"], *PLA Daily*, March 21, 2023. As of June 16, 2023: http://www.81.cn/szb_223187/szbxq/index.html?paperName= jfjb&paperDate=2023-03-21&paperNumber=07&articleid=901476

Huang, Aaron, *Combatting and Defeating Chinese Propaganda and Disinformation: A Case Study of Taiwan's 2020 Elections*, Harvard Kennedy School Belfer Center for Science and International Affairs, July 2020.

Huguet, Alice, Garrett Baker, Laura S. Hamilton, and John F. Pane, *Media Literacy Standards to Counter Truth Decay*, RAND Corporation, RR-A112-12, 2021. As of June 16, 2023: https://www.rand.org/pubs/research_reports/RRA112-12.html

Hwang, Tim, *Deepfakes: A Grounded Threat Assessment*, Center for Security and Emerging Technology, July 2020.

Insikt Group, "Chinese Influence Operations Evolve in Campaigns Targeting Taiwanese Elections, Hong Kong Protests," *Recorded Future* blog, April 29, 2020. As of March 20, 2023: https://www.recordedfuture.com/chinese-influence-operations

International Institute for Strategic Studies, "Large Language Models: Fast Proliferation and Budding International Competition," *Strategic Comments*, Vol. 29, No. 6, March 2023.

Ji Chengfei [纪程飞], Sun Chao [孙超], and Yu Defang [于德芳], "A Preliminary Study of Cyber Psychological Warfare in Informationized Warfare" ["信息化战争中的网络心理战初探"], *Training and Technology* [训练与科技], Vol. 26, No. 6, November 2005.

Jiang, Ben, "Beijing Leads China with the Most Number of AI Firms, as Nation's Capital Pledges Support for Developing ChatGPT-Like Services," *South China Morning Post*, February 14, 2023.

Jiang, Ben, and Coco Feng, "ChatGPT Has Grabbed Headlines but Developing a Chinese Competitor Will Face Censorship, Cost and Data Challenges," *South China Morning Post*, February 20, 2023a.

Jiang, Ben, and Coco Feng, "China's Leadership Wants to Embrace AI Advances but Also Control Risks, as ChatGPT Shocks with Power and Popularity," *South China Morning Post*, April 29, 2023b.

Kaiman, Jonathan, "Free Tibet Exposes Fake Twitter Accounts by China Propagandists," *The Guardian*, July 22, 2014.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," *arXiv*, January 23, 2020.

Knight, Will, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023.

Kuo, Lily, "World's First AI News Anchor Unveiled in China," *The Guardian*, November 8, 2018.

Lajka, Arijeta, "New AI Voice-Cloning Tools 'Add Fuel' to Misinformation Fire," Associated Press, February 10, 2023.

Laris, Michael, "Beijing Blames America for Tiananmen Protests," *Washington Post*, May 31, 1999.

Lee, Yimou, and I-hwa Cheng, "Paid 'News': China Using Taiwan Media to Win Hearts and Minds on Island—Sources," Reuters, August 8, 2019.

Li Bicheng [李弼程], "Model for a System of Online Public Opinion Struggle and Countermeasures" ["网络舆论斗争系统模型与应对策略"], *National Defense Technology* [国防科技], October 2016.

Li Bicheng [李弼程] and Gong Zhenkai [龚振凯], "Affective Text Generation with Hard Constraints" ["硬约束限制的情感文本生成方法研究"], *Application Research of Computers* [计算机应用研究], January 2023.

Li Bicheng [李弼程], Hu Huaping [胡华平], and Xiong Yao [熊尧], "Intelligent Agent Model for Online Public Opinion Guidance" ["网络舆情引导智能代理模型"], *National Defense Technology* [国防科技], June 2019.

Li Bicheng [李弼程], Xiong Yao [熊尧], Huang Tao [黄涛], and Pan Le [潘乐], "Simulation Deduction Model and System Construction for Intelligent Online Public Opinion Guidance" ["网络舆论智能引导仿真推演模型与系统构建"], *National Defense Technology* [国防科技], October 2020.

Li Hongqian [李洪乾] and Liang Xiaobo [梁晓波], "The Problems of China's Defense Language Education Program and Their Countermeasures" ["语言战略化背景下的我国国防语言教育现状及策略研究"], *Journal of Yunmeng* [云梦学刊], Vol. 39, No. 2, 2018.

Li, Qiaoyi, and Eduardo Baptista, "Chinese Organisations Launched 79 AI Large Language Models Since 2020, Report Says," Reuters, May 30, 2023.

Li, Robin [李彦宏], "Baidu Ernie Bot Press Conference" ["百度文心一言新闻发布会"], Baidu Live [百度直播], March 16, 2023. As of June 16, 2023:
https://live.baidu.com/m/media/pclive/pchome/live.html?room_id=8117393980&source=search

Li Weijie [李维杰], "The Influence and Reshaping of the Internet on Western Politics" ["网络对西方政治的影响与重塑"], *China Information Security* [中国信息安全], May 2018.

Liang Xiaobo [梁晓波], "Cognitive Intelligence Language Weapon" ["认知智能语言武器"], *PLA Daily*, December 20, 2019. As of June 16, 2023:
http://www.81.cn/jfjbmap/content/2019-12/20/content_250331.htm

Ling Jiwei [凌纪伟], "Forge Ahead on a New Journey, Entrepreneurs Discuss High-Quality Development, Sun Weijie, CEO of DPTechnology: Create Micro-Scale Scientific Research and Industrial R&D Infrastructure" ["奋进新征程 企业家共论高质量发展, 深势科技CEO孙伟杰：打造微观尺度科学研究和工业研发基础设施"], Xinhua, December 28, 2022. As of June 16, 2023:
http://www.news.cn/tech/20221228/3ae0eb5e0b3e43d9952161992d5a911f/c.html

Liu, Haokun, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel, "Few-Shot Parameter-Efficient Fine-Tuning Is Better and Cheaper Than In-Context Learning," *Advances in Neural Information Processing Systems*, Vol. 35, 2022.

Liu Huiyan [刘惠燕], Xiong Wu [熊武], Wu Xianliang [吴显亮], and Mei Shunliang [梅顺量], "Several Thoughts on Promoting the Construction of Cognitive Domain Operations Equipment for the Omni-Media Environment" ["全媒体环境下推进认知域作战装备发展的几点思考"], *National Defense Technology* [国防科技], Vol. 39, No. 5, October 2018.

Liu Xia [刘 霞], "Technology Industry Trends Looking Ahead to 2023" [展望2023年的技术行业趋势], *Science and Technology Daily* [科技日报] via Xinhua Online [新华网], January 13, 2023. As of June 16, 2023:
http://www.news.cn/tech/20230113/f15ff0ec9ae34a028b95eaafda48a2b7/c.html

Liu Yongdan [刘永丹] and Zhang Yu [张煜], "On Innovation of Military Political Work from the Perspective of Big Data" ["论大数据视域下的军队政治工作创新"], *China Military Science* [中国军事科学], 2016.

Long Yameng [龙亚蒙] and Zhou Yang [周洋], "Research on the Application of Social Bots in Public Opinion Struggle" ["社交机器人在舆论斗争中的应用研究"], *Military Correspondent* [军事记者], 2022.

Luo Yuzhen [罗语嫣], Li Wei [李璜], Wang Ruifa [王瑞发], Lei Wei [雷潇], Liao Dongsheng [廖东升], and Zhu Yingying [朱莹莹], "Characteristics and Key Technologies of the Common Domain for the Cognitive Domain" ["认知域的公域特性及其关键技术"], *National Defense Technology* [国防科技], April 2018.

McGuffie, Kris, and Alex Newhouse, "The Radicalization Risks of GPT-3 and Advanced Neural Language Models," Middlebury Institute of International Studies Center on Terrorism, Extremism, and Counterterrorism, September 14, 2020.

McKune, Sarah, "An Analysis of the International Code of Conduct for Information Security," Citizen Lab, September 28, 2015.

Ministry of Foreign Affairs of the People's Republic of China, "China, Russia and Other Countries Submit the Document of International Code of Conduct for Information Security to the United Nations," September 13, 2011.

Ministry of Foreign Affairs of the People's Republic of China, "International Strategy of Cooperation on Cyberspace," March 1, 2017.

Ministry of Foreign Affairs of the People's Republic of China, "Fact Sheet: U.S. Interference in Hong Kong Affairs and Support for Anti-China, Destabilizing Forces," September 24, 2021a.

Ministry of Foreign Affairs of the People's Republic of China, "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)," December 14, 2021b.

Ministry of Foreign Affairs of the People's Republic of China, "Fact Sheet on the National Endowment for Democracy," May 7, 2022.

Ministry of Foreign Affairs of the People's Republic of China, "US Hegemony and Its Perils," February 20, 2023.

Mostaque, Emad [@EMostaque], "We actually used 256 A100s for this per the model card, 150k hours in total so at market price $600k," Twitter post, August 28, 2022. As of June 16, 2023: https://twitter.com/EMostaque/status/1563870674111832066

Mueller, Robert S. III., *Report on the Investigation into Russian Interference in the 2016 Presidential Election*, Vol. I, U.S. Department of Justice, March 2019.

Murayama, Keiichi, and Mitsuru Obe, "Microsoft President Warns China Becoming Close Rival of ChatGPT," *Nikkei Asia*, April 21, 2023.

Nemr, Christina, and William Gangware, *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*, U.S. Department of State, March 2019.

OpenAI, "Fine-Tuning," webpage, undated. As of June 16, 2023: https://platform.openai.com/docs/guides/fine-tuning

OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," November 30, 2022. As of January 23, 2023: https://openai.com/blog/chatgpt/

Rochefort, Alex, "Regulating Social Media Platforms: A Comparative Policy Analysis," *Communication Law and Policy*, Vol. 25, No. 2, 2020.

"Russian Twitter Political Protests 'Swamped by Spam,'" BBC News, March 8, 2012.

Ryan, Fergus, Ariel Bogle, Nathan Ruser, Albert Zhang, and Daria Impiombato, *Borrowing Mouths to Speak on Xinjiang*, Australian Strategic Policy Institute, December 2021.

Sætra, Henrik Skaug, "Generative AI: Here to Stay, but for Good?" Social Science Research Network, March 22, 2023.

Satariano, Adam, and Paul Mozur, "The People Onscreen Are Fake. The Disinformation Is Real." *New York Times*, February 7, 2023.

Sedova, Katerina, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, *AI and the Future of Disinformation Campaigns, Part 1: The RICHDATA Framework*, Center for Security and Emerging Technology, December 2021a.

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 2021b.

Shan, Shelley, "China Might Use AI to Sow Chaos: NSB," *Taipei Times*, April 27, 2023.

Shen Bilong [沈弼龙], "Military Application of Large Model Technology" ["大模型技术的军事应用"], *PLA Daily*, April 11, 2023. As of June 16, 2023: http://www.81.cn/jfjbmap/content/2023-04/11/content_337361.htm

Shen Zhengzheng [申峥峥] and Shu Zhe [束哲], "Generative AI: How Far Is It from Comprehensive Application in the Military Field" ["生成式人工智能: 距离军事领域全面应用有多远"], *PLA Daily*, April 14, 2023. As of June 16, 2023: http://www.81.cn/yw_208727/16216758.html

Shinn, Noah, Beck Labash, and Ashwin Gopinath, "Reflexion: An Autonomous Agent with Dynamic Memory and Self-Reflection," *arXiv*, March 20, 2023.

Song Yumeng [宋玉萌], "Xinhua News Agency's Intelligent Editorial Department Is Up and Running, Realizing Artificial Intelligence to Reengineer the Whole Process of News Production" ["新华社智能化编辑部建成运行 实现人工智能再造新闻生产全流程"], Xinhua, December 12, 2019. As of June 16, 2023: http://www.xinhuanet.com/politics/2019-12/12/c_1125340864.htm

Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese," in Ricardo Cerri and Ronaldo C. Prati, eds., *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings*, Springer, October 2020.

State Council Information Office of the People's Republic of China, *China's National Defense in the New Era*, July 2019.

Strick, Benjamin, *Analysis of the Pro-China Propaganda Network Targeting International Narratives*, Centre for Information Resilience, 2021.

Toh, Michelle, "Baidu Stock Rebounds After Falling Sharply in Wake of ChatGPT-Style Bot Demo," CNN, March 17, 2023.

U.S. Department of State, Bureau of Arms Control, Verification, and Compliance, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," February 16, 2023.

U.S. Government Accountability Office, *Contested Information Environment: Actions Needed to Strengthen Education and Training for DOD Leaders*, January 2023.

Wang Jinxia [王金霞], "Affective Computing: Achieving Intelligent Interaction Technology" ["情感计算: 成就智能交互技术"], *PLA Daily*, April 8, 2023. As of June 16, 2023:
http://www.81.cn/jfjbmap/content/2022-04/08/content_313202.htm

Wang Yunlong [王云龙] and Zhang Zhiwei [张智伟], "Emerging Communication Technology Revolution and Developmental Trends in Military Struggle for Public Opinion" ["新兴传播技术革命与军事舆论斗争发展趋势"], *China Military Science* [中国军事科学], August 2020.

Weise, Karen, "Microsoft's Bing Chatbot Offers Some Puzzling and Inaccurate Responses," *New York Times*, February 15, 2023.

Weng Chen [翁辰] and Chen Qiaojian [陈俏健], "A Preliminary Study on the Operation Ideas and Methods of the Facebook Social Platform" ["Facebook社交平台运营思路及方法初探"], *Public Communication of Science and Technology* [科技传播], July 2018.

West, Darrell M., "How to Combat Fake News and Disinformation," Brookings Institution, December 18, 2017.

Wu Jieming [吴杰明] and Liu Zhifu [刘志富], *An Introduction to Public Opinion Warfare, Psychological Warfare, and Legal Warfare* [舆论战心理战法律战概论], National Defense University Press, 2014.

Xinhua, "During the 12th Collective Study of the Political Bureau of the CCP Central Committee, Xi Jinping Emphasized Promoting In-Depth Development of Media Integration and Consolidating the Common Ideological Foundation of the Whole Party and the People of the Country" ["习近平在中共中央政治局第十二次集体学习时强调 推动媒体融合向纵深发展 巩固全党全国人民共同思想基础"], January 25, 2019. As of June 16, 2023:
https://www.12371.cn/2019/01/25/ARTI1548411219417372.shtml

Xinhua, "Xi Jinping Stressed During the 30th Collective Study of the Political Bureau of the Communist Party of China Central Committee the Need to Strengthen and Improve International Communication Work to Showcase a Real, Three-Dimensional and Comprehensive China" ["习近平在中共中央政治局第三十次集体学习时强调 加强和改进国际传播工作 展示真实立体全面的中国"], June 1, 2021, translation via Adam Ni and Yun Jiang, "Xi on External Propaganda and Discursive Power," China Neican, June 4, 2021. As of June 16, 2023:
https://www.neican.org/xi-jinping-on-external-propaganda/

Yang, Charlotte, "ChatGPT Fervor Is so Hot That Chinese Firms Call for Caution," Bloomberg, February 13, 2023.

Yang, Zeyi, "Chinese Creators Use Midjourney's AI to Generate Retro Urban 'Photography,'" *MIT Technology Review*, March 29, 2023.

Yang Chengping [杨成平] and He Wei [何秧], "The Main Contradictions and Countermeasures in Wartime Political Work" ["战时政治工作面临的主要矛盾及对策"], *Journal of Political Work* [政工学刊], November 2007.

Yuan, Li, "Why China Didn't Invent ChatGPT," *New York Times*, February 17, 2023.

Zeng Huafeng [曾华锋] and Shi Haiming [石海明], *Command of the Mind: The Rules of War and National Security Strategy in the Global Media Age* [制脑权: 全球媒体时代的战争法则与国家安全战略], Academy of Military Science Press, 2014.

Zhang, Albee, and Brenda Goh, "Factbox: Chinese Firms Working on ChatGPT-Style Technology," Reuters, February 22, 2023.

Zhang, Albert, Tilla Hoja, and Jasmine Latimore, *Gaming Public Opinion: The CCP's Increasingly Sophisticated Cyber-Enabled Influence Operations*, Australian Strategic Policy Institute, April 2023.

Zhang Tong, "China's Internet Users Finding Creative Uses for ChatGPT," *South China Morning Post*, January 12, 2023.

Zhang Yan [张艳] and Wang Lingshuo [王凌硕], "In This Year of Technological Innovation, Review the Technological Changes That Have Brought Us Many Surprises" ["科技创新这一年, 回顾那些带给我们诸多惊喜的科技之变"], *People's Daily*, December 30, 2022. As of June 16, 2023:
http://www.mod.gov.cn/education/2022-12/30/content_4929518.htm

Zhao Shuang [赵爽] and Feng Haochen [冯浩宸], "Evaluation and Analysis of the Development and Influence of 'Botnets'" ["机器人水军发展与影响评析"], *China Information Security* [中国信息安全], November 2017.

Zhao, Suisheng, ed., *Chinese Authoritarianism in the Information Age: Internet, Media, and Public Opinion*, Routledge, 2018.

Zheng, Sarah, "China's Answers to ChatGPT Have a Censorship Problem," Bloomberg, May 2, 2023.

Zhou, Cissy, "China Tells Big Tech Companies Not to Offer ChatGPT Services," *Nikkei Asia*, February 22, 2023.

About This Perspective

This Perspective explores the implications that generative artificial intelligence (AI) might have for social media manipulation. We argue that the advent of generative AI presents a revolutionary improvement for the social media manipulation process, including content generation and content delivery, and that this heralds a new era. Not everything will change, however, as it is not clear that existing generative AI excels at tasks such as social media campaign design or campaign assessment.

This Perspective begins with an overview of the generational shift in social media manipulation presented by generative AI and an overview of generative AI. We then address the potential threat of generative AI for social media manipulation, including how generative AI will change (and may not change) common social media manipulation tactics and how, in particular, this might affect China's approach to social media manipulation. We provide an overview of China's indigenous generative AI capabilities, explore Chinese military writings that provide insights into how China might leverage these new capabilities, and consider what this might mean for future Chinese efforts against Taiwan as an illustrative case study for this new risk. We also address the likely limitations. We conclude with recommendations for technical, policy, and diplomatic mitigations by U.S. government and industry. We argue that any mitigation strategy must account for generative AI being ubiquitous and unregulated globally.

This research was conducted within the International Security and Defense Program of the RAND National Security Research Division (NSRD), which operates the RAND National Defense Research Institute (NDRI), a federally funded research and development center (FFRDC) sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense intelligence enterprise.

For more information on the RAND International Security and Defense Program, see www.rand.org/nsrd/isdp or contact the director (contact information is provided on the webpage).

The world may remember 2022 as *the year of generative artificial intelligence (AI)*: the year that large language models (LLMs), such as OpenAI's GPT-3, and text-to-image models, such as Stable Diffusion, marked a sea change in the potential for social media manipulation. LLMs that have been optimized for conversation (such as ChatGPT) can generate naturalistic, human-sounding text content at scale, while open-source text-to-image models can generate photorealistic images of anything (real or imagined) and can do so at scale. Using existing technology, U.S. adversaries could build digital infrastructure to manufacture realistic but inauthentic (fake) content that could fuel similarly realistic but inauthentic online human *personae*: accounts on Twitter, Reddit, or Facebook that seem real but are synthetic constructs, fueled by generative AI and advancing narratives that serve the interests of those governments.

In this Perspective, the authors argue that the emergence of ubiquitous, powerful generative AI poses a potential national security threat in terms of the risk of misuse by U.S. adversaries (in particular, for social media manipulation) that the U.S. government and broader technology and policy community should proactively address now. Although the authors focus on China and its People's Liberation Army as an illustrative example of the potential threat, a variety of actors could use generative AI for social media manipulation, including technically sophisticated nonstate actors (domestic as well as foreign). The capabilities and threats discussed in this Perspective are likely also relevant to other actors, such as Russia and Iran, that have already engaged in social media manipulation.

RAND

NATIONAL SECURITY RESEARCH DIVISION

www.rand.org