THE IMPACT OF GENERATIVE AI ON UPCOMING ELECTIONS: DISINFORMATION AND OTHER MALIGN INFLUENCE OPERATIONS COMING OUR WAY AT SCALE

There are a number of threats posed by large language models (*see generally*, my earlier briefing note, available here) and, given the ways in which the technology builds on itself, addressing the myriad threats is urgent. That said, one threat that needs to be understood and addressed as a matter of particular urgency is the potential, in the words of OpenAI CEO Sam Altman, for the technology to provide "interactive disinformation" in the run up to the US elections. Ditto for the United Kingdom, as the next general election must be held by January 2025, and most likely will be scheduled for October 2024.

Interactive disinformation, in short, refers to the ability of ChatGPT and other large language model technologies, based on their fast-evolving abilities to predict human reactions, to manipulate and persuade voters on a one-on-one basis, at scale. This capability in the context of upcoming elections, prompted Altman, as part of his testimony before the Senate Judiciary Committee, Subcommittee on Privacy, Technology, and the Law on May 16 (transcript available here), to characterize the threat of interactive disinformation as a "significant area of concern" and one of his areas of "greatest concern." Altman's comment was in response to a question posed by Senator Josh Hawley about the ability of large language models to predict public opinion and the potential for a range of actors, including domestic campaigns and foreign actors, to abuse survey information, in the context of elections, to elicit desired responses.[1]

Regulators and policymakers across the globe are intensely focused on generative AI, but legislative processes may prove to be too unwieldy and too slow to staunch the malign uses of the models in upcoming elections. That said, as generative AI requires a whole-of-society response, there are roles for many others.

As noted in my two prior briefing notes on generative AI, there are countless potential positive contributions of the technology, including in the context of political campaigning. Those contributions should not be viewed as diminished by the potential. malign uses.

Malign Influence Operations

Following Russian efforts to influence the 2016 Brexit referendum (yet to be fully analyzed) and to influence the US elections (analyzed at length, for example, in the Report on the Investigation Into Russia Interference in The 2016 Presidential Election), I asked the question: how was it that the Internet Research Agency and other malign Russian actors managed to have a better grasp than we did of the percolating culture wars in the United Kingdom and, particularly, in the United States? Was it sheer luck, was it

[1] Senator Hawley's question was based on conclusions set out in a March 2023 paper written by Eric Chu, Jacob Andreas, Stephen Ansolabehere and Deb Roy, "Language models trained on media diets can predict public opinion".

the product of on-the-ground analysis, was it AI? In 2015-2016, it was not AI. In 2023, generative AI is fast becoming the feared source of malign influence – both domestic and foreign.

NYU Professor Emeritus Gary Marcus, in his testimony at the May 16 hearing, sounded an ominous note: "Fundamentally, these new systems are going to be destabilizing. They can and will create persuasive lies at a scale humanity has never seen before. Outsiders will use them to affect our elections, insiders to manipulate our markets and our political systems. Democracy itself will be threatened. Chatbots will also clandestinely shape our opinions, potentially exceeding what social media can do."

A joint effort undertaken by OpenAI, Georgetown University's Center for Security and Emerging Technologies (CSET) and the Stanford Internet Observatory analyzed the potential impact of generative AI on three aspects of influence operations – actors undertaking the campaigns, deceptive behaviors leveraged as tactics, and the content itself. The analysis was driven by the potential of large language models to mimic human-generated written content, at scale and at low cost, while becoming easier to use, more adept at producing outputs without obvious errors and more cost-efficient. (The focus on large language models was intentionally narrow; the potential uses of videos, images and multimodal models (*e.g.,* deepfakes) were beyond the scope of the study because text is particularly difficult to distinguish as AI-generated, and access to large language models is spreading much more quickly.)

The study adopted as its definition of influence operations, "covert or deceptive efforts to influence the opinions of a target audience." As such, these operations can activate people who hold a particular belief, can persuade an audience of a particular viewpoint and/or can distract a target audience. While the content often can be characterized as disinformation, influence operations are a broader category of malign activity.[2] These operations can be foreign campaigns (often using digital agents of influence to appear to be local) or domestic campaigns. These operations can have the desired impact based on specific content or focus (by persuading a target of a view or reinforcing one, or by distracting a target from finding other ideas or simply causing a target to shut down any further inquiry). They can also erode trust in the overall information ecosystem. The 2016-2022 vintage influence operations broadly were constrained by resource limitations, poorly constructed or reasoned arguments, and ease of detectability. That is changing in real time.

As set out in the January 2023 Report on Emerging Threats and Potential Mitigations, the OpenAI-Stanford-CSET study concluded that large language models:

[2] The OpenAI-Stanford-CEST Report notes that researchers tend to focus on actors, behavior *and* content because one or more components of malign influence operations may, in fact, be authentic. Authentic content may be disseminated by inauthentic actors or authentic actors may use inauthentic automation.

- by replacing or augmenting human writers in the content-generation process, will drive down the cost of generating propaganda, enabling more malign actors to wage influence operations;
- through automated text generation, will create opportunities for propagandists-for-hire and will make propaganda campaigns easier to undertake;
- will enable cross-platform testing, potentially allowing actors to test audience reaction faster and at greater scale, and could overwhelm government public comment processes;
- have the potential to make disinformation more credible and persuasive by compensating for malign actors' lack of linguistic or cultural knowledge of targets and make disinformation harder to discover by in effect replacing copy-paste efforts with linguistically distinct messaging; and
- have the potential to enable novel tactics such as dynamic, personalized, real-time content generation. Dynamic content generation could also be deployed via chatbots, resulting in fake interactive social media persons, interactive email messaging and fake support chatbots. The Report notes, based on research around COVID-19 vaccines, that fake chats could be particularly pernicious vectors of influence.

Deepfakes

Concerns about the confluence of technology, disinformation and elections predate the emergence of ChatGPT and other large language models. For example, in 2018, writing in Foreign Affairs, Robert Chesney and Danielle Citron ("Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics") suggested that as "deepfake technology develops and spreads, the current disinformation wars may soon look like the propaganda equivalent of the era of swords and shields." They noted that, in the context of severe polarization over basic facts and social media platforms facilitating "information cascades," whereby users forward information shared by others without verifying accuracy and in the process enhancing credibility, deepfakes will have "potentially explosive implications for politics."

While deepfakes can have an impact across the entire political landscape, as Catherine Powell and Alexandra Dent noted in a post for the Council on Foreign Relations ("Artificial Intelligence Enters the Political Arena" - May 2023), emerging technologies tend to be used most effectively against women, people of color and members of the LGBTQI+ community running for political office. They also give rise to the risk of what Chesney and Citron called in 2018 the "liars' dividend," allowing politicians to evade responsibility for real videos, audio or photos by claiming they were AI-generated or a deepfake. We are on the brink of a world where everything could be fake.

Flooding the Zone

Just as generative AI has the power to transform the written word, so too can it turbocharge weaponization of imagines, audio and videos. And in both cases, the threats

are, as articulated by Noémi Bontridder and Yves Poullet in their post, "The role of artificial intelligence in disinformation," two-fold: they can be leveraged to create realistic content to manipulate at scale and they can amplify the spread of the malicious content.

While commentators have catalogued a series of shortcomings of large language models, including hallucinations, the potential for these models to sow further discord and distrust should by no means be underestimated. The threats are particularly concerning because of the tendency of people to believe falsehoods that are widely in circulation. Malign actors can exploit this via "astroturfing" – flooding the zone with misinformation and disinformation. As Nikolas Guggenberger and Peter N. Salib noted in their Lawfare post, "From Fake News to Fake Views: New Challenges Posed by ChatGPT-Like AI," as generative AI allows chatbots to be "indistinguishable from humans, humans will start to question the identity of everyone online, especially those with whom they disagree."

Addressing the Threats

The OpenAI-Stanford-CSET Report calls for a whole-of-society response, including collaboration between AI providers and social media platforms, and strong norms within the AI community around release of large language models or the training methods used to develop them. In an interview with Cointelegraph, Associate Professor Trish McCluskey (Deakin University) also calls for a whole-of-society approach, involving "government regulation, self-regulation by tech companies, international cooperation, public education, technological solutions, media literacy and ongoing research." (*See* "Forget Cambridge Analytica - Here's how AI could threaten elections".)

Altman, in his May 16 testimony, posited that "regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models." He testified that he welcomes regulation of generative AI in the form of:

- disclosure requirements ("people need to know if they're talking to an AI");
- licensing requirements; and
- internal testing as well as external testing (independent audits) prior to release of models (to ensure the models cannot "self-replicate" or more colorfully "exfiltrate into the wild"), with evaluation results made publicly available.

Altman also called for the establishment of a new regulatory agency to license activities above a "certain scale of capabilities" and to ensure compliance with safety standards. Marcus though cautioned about the risk of "regulatory capture" – a greenwashing-style effect making it appear as if regulation is working, when in fact it is not. Altman and Marcus also endorsed the need for international oversight, with Altman suggesting a body modelled on the Vienna-based International Atomic Energy Agency (Marcus likewise has endorsed an IAEA-like agency, in a recent article, "The world needs an

international agency for artificial intelligence, say two AI experts, co-authored with Anka Reuel in The Economist.[3]

IBM Chief Privacy and Trust Officer Christina Montgomery called for risk-weighted precision regulation (*i.e.*, different rules for different risks), including, in the case of elections, rules around disclosure of the data being used to train the models and the performance of the algorithms being used "in that context."

Marcus referred to what he views as the central scientific issue in AI, namely that we do not know how to build a system that understands the full scale of harm. He called for "nutrition labels" – how do the models generalize? What goes into the systems? He noted that the risk of manipulation by the models, even if not intentional, is very real and fears that interacting with opinionated language models can change user views. To address these issues, he posits that we need to understand what the models are trained on – this proposal was not part of Altman's proposed guardrails. Transparency about data requires not only access, but third-party access by independent experts.

Marcus also noted the limitations. We do not have the tools today to detect and label misinformation. Senator Alex Padilla raised a separate concern around "repeating social media's failures" in investing only in English language tools.

Yesterday, in Washington, D.C., Microsoft Vice Chair and President Brad Smith joined the conversation in a speech coinciding with the release of Microsoft's five-point plan, "Governing AI: A Blueprint for the Future". The five-point plan calls for:

- new government-led AI safety frameworks;
- safety breaks for AI systems that control critical infrastructure;
- a broader legal and regulatory framework, including imposing regulatory responsibilities on the industry, creating a new government agency to regulate "highly capable AI foundation models," and imposing labelling requirements akin to bank KYC requirements;
- transparency and academic/nonprofit community access to AI resources; and
- new public-private partnerships.

In response to the "challenge of the 21st century," while "companies need to step up," "government needs to move faster."

---

[3] The world is in a very different place than it was in 1957 when the IAEA was formed in response to fears about the development and spread of nuclear weapons. Today, the European Union is moving forward (*see* my previous briefing note, available here) with a legal regime to regulate AI (with the European Parliament weighing in with stricter transparency requirements aimed at large language models), prompting Altman to warn (according to reporting in Time Magazine), while speaking at a conference in London earlier this week, that these regulatory efforts could lead OpenAI to cease operating in the European Union.

Regulation?

At the May 16 hearing, Judiciary Subcommittee on Privacy, Technology, and the Law Chair Senator Richard Blumenthal was clear:

> "AI companies ought to be required to test their systems, disclose known risks, and allow independent researcher access. We can establish scorecards and nutrition labels to encourage competition based on safety and trustworthiness, limitations on use. There are places where the risk of AI is so extreme that we ought to restrict or even ban their use, especially when it comes to commercial invasions of privacy for profit and decisions that affect people's livelihoods. And of course, accountability, reliability. When AI companies and their clients cause harm, they should be held liable."

To illustrate the potential harm, the introduction to his remarks – both the content and the delivery – were created by AI. Blumenthal was blunt in his prescription: "We need to maximize the good over the bad. Congress has a choice. Now. We had the same choice when we faced social media. We failed to seize that moment. Now we have the obligation to do it on AI before the threats and the risks become real."

The track record these days when it comes to Congressional regulation and oversight of big tech, hampered by fundamental divisions between Democratic and Republican lawmakers, does not bode well. There are clear paths to address the amplification of disinformation by social media platforms – the EU Digital Services Act is the best, and only, example to date – but US efforts are stalled, and self-regulation is generally deemed to have been woefully inadequate. Obviously too, standing in the way of regulation is the China question and a disinclination to hamper US innovation, at the expense of maintaining a competitive edge.

That said, earlier this month, Congresswoman Yvette D. Clarke introduced the REAL Political Advertisements Act, which would expand current disclosure requirements for campaign ads if generative AI were used to generate any videos or images in the ad. Senators Michael Bennet and Peter Welsh have proposed an updated version of legislation calling for the establishment of a Federal Digital Platform Commission. And earlier this week, the White House Office of Science and Technology Policy published a request for information on harnessing the benefits of, and mitigating the risks of, AI, as part of the national AI strategy that is under development.

Concluding Thoughts

Generative AI has the potential to transform every facet of life as we know it – in hugely beneficial ways and in hugely harmful ways. Politics will be no exception.

Campaign staffs will be able to respond in real time to developments – just as the RNC did when it aired an AI-generated dystopian video immediately following President

Biden's re-election announcement.[4] Generative AI has the potential to enable candidates to generate campaign materials and responses in a matter of minutes, at low cost, end-running advisors and consultants. Campaigns will then be able to use generative AI to fine-tune targeted messages and analyze vastly increased datasets, improving targeting of voters for GoTV purposes and donors for fundraising purposes.

These same tools will be available across the public space, in effect enabling almost anyone to inject AI-generated messages into the social media sphere.

And then we layer in the malign uses detailed above.

AI experts (as noted above) speak of the need for a whole-of-society of response, which by definition has multiple components. We should hope that as many components as possible will be activated, starting with creating public awareness. Just as there were countless efforts around both the 2020 and 2022 elections to educate the public (*see, e.g.*, the Election Security Rumor vs. Reality page and We're in This Together fact sheet on the Cybersecurity and Infrastructure Security Agency website and the #TrustedInfo page on the National Association of Secretaries of State website), so too must a concerted effort be made to create a resilient, informed electorate.

This briefing note is intended to help kickstart those efforts.

* * *

Mark S. Bergman
7Pillars Global Insights, LLC
Washington, D.C.
May 26, 2023

No portion of this briefing note was prepared using ChatGPT.

[4] The RNC did include a disclaimer on its webpage acknowledging the use of AI, but malign foreign and other actors will not. Calling for disclaimers more broadly needs to take account of how easy it is for a disclaimer to fall away. The RNC disclaimer does not accompany the YouTube versions of the video, accessible directly or even via the RNC webpage.