AI Forum statement by Michael Chertoff submitted 11/6/2023

Introduction

Hostile state actors and their surrogates are increasing their use of tradecraft that leverages connected technology to influence and interfere with elections. In past election cycles, we have seen: (1) theft of confidential information; (2) false attacks on candidates; (3) misleading statements on how influential individuals plan to vote; (4) attacking the legitimacy of the elections process. These acts are bad but not new—for example, the Presidential election of 1800 between John Adams and Thomas Jefferson was notoriously toxic and full of false accusations of, among other things, criminal activity and sexual misconduct – however, advances in artificial intelligence could be used by threat actors to magnify these dynamics by:

(1) Powering increasingly sophisticated cyber threat techniques that increase an adversary's ability to penetrate election-related networks (i.e., voter registration, as well as campaign and news media systems, etc.).
(2) Creating fake content that appears highly realistic.
(3) Manipulating algorithmic systems to influence or disenfranchise voters.

Of particular concern is how these techniques may be combined – for example, using a cyber intrusion to enable strategic access to a system, populate that system with disinformation and then disseminate it in a highly-targeted way.

Over the next twelve or so months, pivotal elections will be held not just in the United States, but in several other major democracies, such as Taiwan (January 2024), Indonesia (February 2024), South Africa (June 2024), Mexico (July 2024), plus India and potentially the United Kingdom. It is critical that we illuminate AI-related election integrity risks so that we can address and mitigate them before it is too late.

Historical Context: Malicious Use of Technology to Influence Elections

In the United States, authorities have disclosed that several U.S. states were the target of Russian government cyber actors seeking vulnerabilities and access to U.S. election infrastructure during preparations for the 2016 presidential elections. Likewise, emails of political leaders were compromised, with the contents being selectively leaked. In 2018, several Russian organizations and more than one dozen Russian nationals were indicted for conspiracy to defraud the United States through activities intended to influence the 2016 presidential election.

This tradecraft is not limited to Russia. Two Iranian nationals were also charged in 2021 with efforts to influence the **2020 U.S. Presidential election** through a combined cyber threat (exfiltrating voting records from a state election website) and disinformation campaign (including both (a) false-flag operations claiming to be from the Proud Boys warning of Democratic National Committee attempts to exploit vulnerabilities in state election systems to submit fake ballots and (b) voter intimidation emails that threatened the recipients with physical injury if they did not change their party affiliation and vote for President Trump).

Targets are not limited to the United States: the campaign of then-French Presidential candidate Emmanuel Macron was hacked on the eve of the 2017 presidential election in France where nine gigabytes of data were strategically released just hours before the French mandated media blackout. Moreover, in 2019, a European Commission report revealed evidence of a "continued and sustained disinformation activity by Russian sources aiming to suppress turnout and influence voter preferences," which caused the EU to implement several EU-wide mitigation measures, including the EU Rapid Alert System.

What's Different About AI

Artificial Intelligence offers great promise across many sectors, ranging from healthcare to security, and yet it also carries with it potentially enormous risks, particularly with the advent of Generative AI, which uses existing training data not just to answer questions but to "generate" new content. President Biden's October 30 Executive Order launches a whole-of-government approach to mitigate AI risks as well as promote AI innovation and the government's responsible use of AI while acknowledging the need for bipartisan legislation to support implementation. AI-related risks[1] tend to fall into three categories: (1) the use of AI technologies as an instrumentality of threat activity, (2) the malicious targeting of AI systems themselves; and (3) unintended consequences associated with innocent use of AI technologies that can have profound political, economic, ethical and other potentially destabilizing effects. All three of these categories figure into election integrity risks.

AI as an Instrumentality of Cyber Adversaries, Disinformation, and Influence

Generative AI can enable threat actors with increasingly powerful tools for cyber intrusions, fraud, and disinformation.

Cyber Intrusions. In the case of cyber intrusions, core election-related threat objectives may include:

- Compromise the confidentiality of sensitive political communications for later use in influence and interference efforts;
- Compromise of the confidentiality of voting infrastructure for intelligence or hostile contingency planning;
- Compromise the integrity of election-day media reporting to impact voter turnout; and
- Compromise the integrity or availability of core election infrastructure.

China and Russia both have agendas that can be advanced through such cyber intrusions. Setting aside potential candidate preferences, creating a level of chaos or inciting mistrust supports both

[1] In this statement, I am principally focused on risks from the use of Generative AI (GAI – e.g., ChatGPT, MidJourney, Stable Diffusion, Dall-E, etc.) and other narrow AI systems such as expert learning models. For the purposes of this statement, I am not addressing risks from Artificial General Intelligence, whose development is rather further in the future.

regimes' positions that democracies are messy and that their systems are superior and that they are justified in acting against democratic protestors for example in Hong Kong.

AI can be used as an instrumentality to facilitate increasingly realistic social engineering campaigns – not only in text or image form but increasingly in voice-related means as well. While not election-related, technology company Retool described in August 2023 how deepfake techniques were used to gain unauthorized access to the company: "The caller claimed to be one of the members of the IT team, and deep faked our employee's actual voice. The voice was familiar with the floor plan of the office, coworkers, and internal processes of the company." Researchers have also demonstrated how platforms like ChatGPT could be used to create polymorphic malware (which mutates to change its code while retaining its core function). These techniques demonstrate the ability of AI to rapidly change tactics to obfuscate malintent from monitoring systems, including those intended to identify illegal actions affecting elections.

Identity Subversion and Voter Fraud. Identity verification is a critical step to preventing and detecting fraud in voter registration processes, and states take varying approaches to the level of identity assurance built into actions like initial registrations and address updates. While not directly voter registration-related, government agencies are starting to see the use of AI technologies to enable fraud in government benefits programs. In July 2023, the Social Security Administration reported that Office of Inspector General agents "discovered that an AI powered "chatbot" was used to impersonate beneficiaries and contact customer service representatives to divert monthly benefit payments to spurious accounts."

Disinformation. In the case of media manipulation that occurs, core election-related objectives may include:

- Disparagement of a candidate or political party, or promotion of a favored candidate;
- Undermining of public confidence in the integrity of elections;
- Dissuading voter turnout; and
- Incitement of physically disruptive actions by disgruntled groups.

Generative AI can also potentially be used as a tool to impersonate an individual or more generally to spread false or misleading information to erode trust. In March 2022, a deepfake video was released of Ukrainian President Volodymyr Zelensky appearing to announce a surrender to Russia. Social media platforms can be a vehicle to spread disinformation at scale, when minimal identity verification mechanisms are in place as regards account holders. There are societal benefits to minimal identity verification as a means to protect anonymity and foster free speech online, but this also means that little transparency exists on the identity of accounts in question.

AI & Use of Voter Messaging Algorithms

For years, political campaigns have sought to conduct targeted messaging campaigns to influence voters. This is a legitimate aspect of political discourse, but the techniques can sometimes be unscrupulous. In 2018, the New York Times broke the story on how Cambridge

Analytica had improperly harvested information from 50 million Facebook users and used that information to help the Trump campaign launch targeted political messaging through psychographic modeling techniques.

The fear is that AI large language model technologies could be utilized to analyze large personal data sets and thereby enable highly targeted messaging with statements that are both persuasive and false. As described in an April 2023 Foreign Affairs article, language models, dubbed "personalized propaganda," have already been trained to persuade game players to partner with them in a game, and they could be readily trained to persuade people to take actions with real world effect, to include changing candidate preferences due to a negative image or the perception that the candidate cannot win.

Issues with Legitimate AI Use

Algorithmic techniques are also used by some election officials to purge voter rolls and for signature verification in absentee ballot scenarios. Such algorithms serve important fraud-prevention imperatives, but (without guardrails) they also have the potential to disenfranchise voters.

What Needs to Be Done

Here is an outline for how policy makers can begin to resolve these risks.

- **Defending Against Election-focused Cyber Intrusions.** This category tends to fall more squarely in the domain of security, fraud, and crisis management teams. In such cases, AI is in the hands of the adversary, and managing such risks requires a threat-informed defense where threat intelligence is used to understand how weaponized AI can weaken the effectiveness of security technologies – for example email filtering or endpoint protection systems – and what countermeasures need to be implemented to address them (e.g., reputational analysis, more user-centric behavioral analytics). For small to midsized organizations without their own threat intelligence teams, like many local government agencies responsible for administering elections, resilience will increasingly depend on cloud service providers using their telemetry to anticipate and withstand AI-enabled attacks on behalf of their customers.

- **Preventing and Disclosing Deepfakes and Disinformation:** Several steps can help address the use of deepfakes in disinformation campaigns:

    - Provenance and/or watermarking systems can play a potentially significant role to determine if a particular piece of content was legitimately created.
    - Proactive disclosures by campaigns using AI to simulate images can provide transparency on such use, and have been the topic of recent public debate related to the use of Federal Election Commission authorities to require such disclosures.

  - Proactive measures to prevent the use of deepfakes to impersonate voters. Identity verification can be an important practice both in combating voter fraud and deterring threat actors seeking to join social media platforms. That said, Congress should examine both the availability of fraud-resistant approaches to identity verification, and the pros/cons of incentivizing or requiring their use in appropriate settings. For voter registration transactions, "liveness detection" tools can help validate that the citizen is a living person. Further, identity documents should now be verified against a trusted system of record, such as the issuing driver license database or validated passport database. The Federal Trade Commission is also warning companies of the potential legal peril they face when releasing AI tools that could be used for fraud or other harm.

- **Disclosures on AI-Powered Messaging.** As noted above, proactive disclosures can also provide transparency where machine learning algorithms and large language models are utilized to develop highly targeted messaging campaigns. While targeted messaging has long been an element of the modern political campaign, the ability to use AI to micro message down to the individual voter presents new risks. Privacy impact assessments for any holders of bulk PII should also disclose whether PII data holdings are used for AI-powered analysis – either directly or through sharing with a third party. We should also consider guardrails around the bulk use of personally identifiable pattern-of-life data for political messaging or campaigning without user consent. This could involve restrictions on Platforms, Campaigns and Political Action Committees (PACs).

- **Mitigating Unintended Consequences of Legitimate Algorithmic Use.** On decision-making, humans should always be in the loop as the final adjudicator for any important AI-supported decision-making.

- **Deep Fake Detection.** The Government should fund research into technologies that can quickly and accurately detect deep fakes. To date, this has been something of a game of cat and mouse, with detection capabilities being able to keep up with AI-powered image and video generation capabilities. Rapid advancements in AI technology, however, threaten to upset this balance as the flaws in deep fakes become more minute. Proposed legislation against deep fakes, such as the "NO FAKES Act," may be able to help, but will not deter determined foreign state actors. International cooperation and the setting of international norms will be important in deterring state-sponsored use of advanced deep fakes.

- **Enablement of Trust and Safety Programs.** Larger social media platforms maintain community standards and Trust and Safety programs to detect and respond to harmful content uploaded to their platforms. These programs are a critical first line of defense against the spread of disinformation and deepfakes through the underlying platforms, and their current maturity varies across platforms. The European Union's recently enacted Digital Service Act imposes a number of obligations to mitigate and respond to harmful content, with additional requirements for Very Large Online Platforms. Generative AI platform providers like Google have also released prohibited use policies, which are enforced through business logic on the platforms themselves, but threat actors are already using prompt engineering techniques to bypass such policies.

Disclosure: For the sake of transparency, note that Michael Chertoff is involved in AI and election integrity efforts, including the Transatlantic Commission on Election Integrity and the American Bar Association's AI Task Force. He also serves as an advisor to Truepic, and the Chertoff Group works with companies interested in AI.