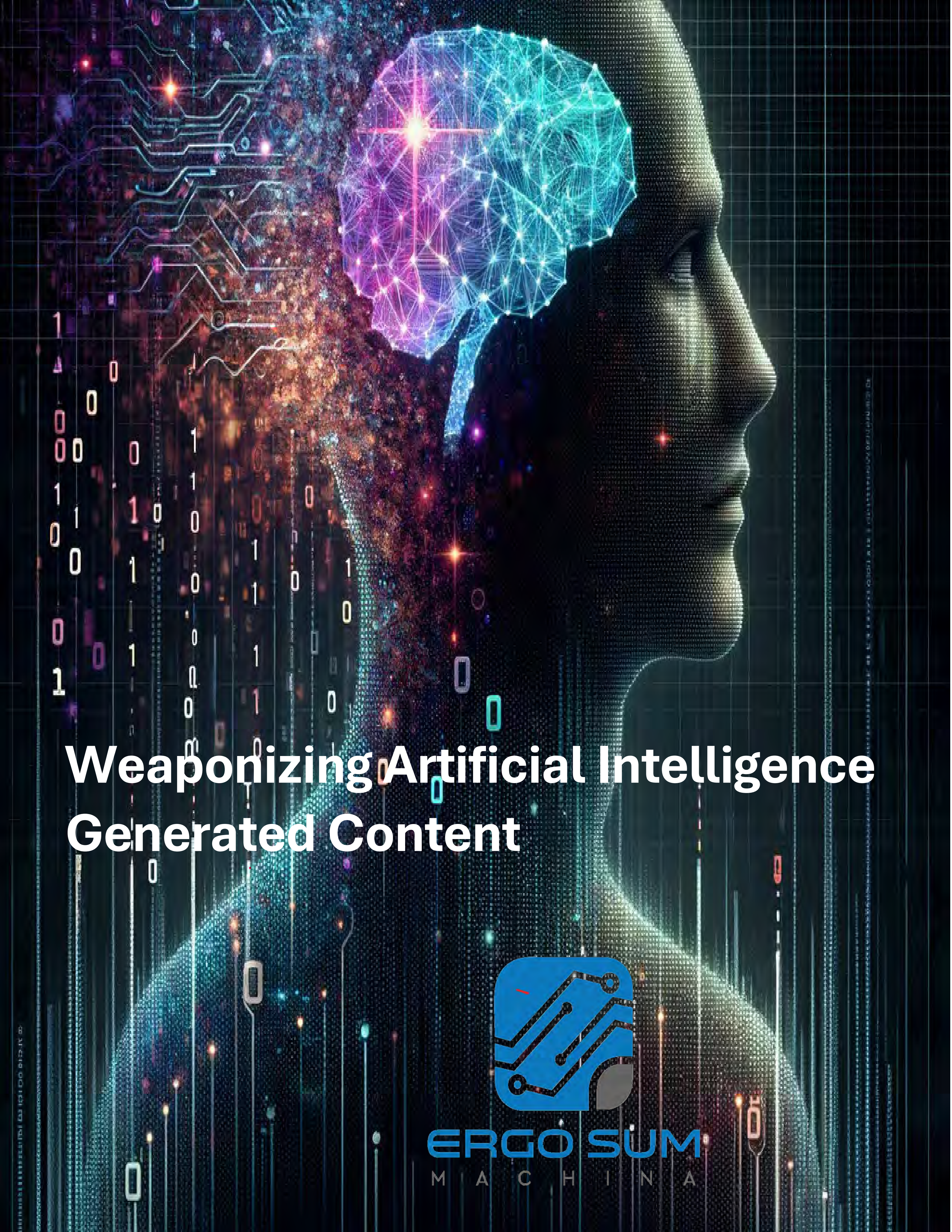


Weaponizing Artificial Intelligence Generated Content

ERGO SUM

MACHINA

(This page intentionally left blank)

Weaponizing Artificial Intelligence Generated Content

By
Team Ergo Sum Machina
Mr. Tom Jackson
COL Robert Richardson
LTC Katherine Ogletree
LTC Charles Moss
CDR Robert Liberato

United States Army War College

Class of 2024

The views expressed herein are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Army, Department of Defense, or the U.S. Government. The U.S. Army War College is accredited by the Commission on Higher Education of the Middle States Association of Colleges and Schools, an institutional accrediting agency recognized by the U.S. Secretary of Education and the Council for Higher Education Accreditation.

(This page intentionally left blank)

About This Document

The members of Team Ergo Sum Machina produced this collective strategic research project as one of the prerequisites for completing the Masters of Strategic Studies program at the College of the United States Army War College (USAWC). The research, analysis, and production of this report were conducted from October 2023 through April 2024 as part of the Army Futures Seminar for Academic Year 2024.

Requirements

This report answers a strategic question posted by LTG Laura Potter, Deputy Chief of Staff, G2, based on the available open-source information and interviews with subject-matter experts.

How might future adversaries use emerging Artificial Intelligence Generated Content (AIGC) tools, techniques, and processes for deception and manipulation of the psycho-cognitive domain (PCD) through 2033?

- What are the likely effects and implications for the use of emerging AIGC tools, techniques, and processes?
- What new detection, attribution, and countermeasure processes will need to be developed to counter the threat of AIGC?

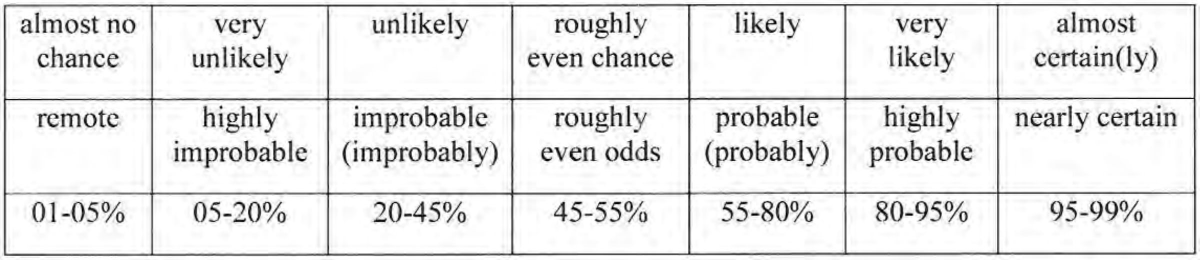
The team's findings were produced in multiple mediums, including a digital PDF version (primary), digital online mind map, and soft-bound book format. Multiple methodologies were used to determine key findings and convergences, including interviews with subject matter experts, scholarly publications, open-source reporting, and the nominal group technique.

Analytic Confidence

The overall estimate was made with moderate analytical confidence. The questions asked were complex, while the timeline was relatively short because of the competing academic requirements of the USAWC core curriculum. Source reliability and corroboration were predominantly moderate to high. However, the analysts were not subject matter experts and worked both individually and collaboratively to research and answer the questions. They utilized a combination of structured analytic techniques, including nominal group technique and network analysis. The Ergo Sum Machina team also evaluated analytic confidence using Peterson analytic confidence factors coupled with Friedman Corollaries (see Annex B).

Words of Estimated Probability

The research team used Intelligence Community Directive (ICD) 203 as their guide for determining their Words of Estimative Probability (WEP) (see Annex C) for expressions of likelihood or probability, an analytic product must use one of the following sets of terms for determining the likely applications across the continuum of AIGC through 2033.



almost no chance	very unlikely	unlikely	roughly even chance	likely	very likely	almost certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01-05%	05-20%	20-45%	45-55%	55-80%	80-95%	95-99%

Source Reliability

Source reliability is noted at the end of each citation as low (L), moderate (M), or high (H). Citations are hyperlinked to the source. Source reliability is determined using Standard Primary Credibility Scale (see Annex D) and the Trust Scale and Website Evaluation Worksheet (see Annex E). Sourced figures and photos embedded in the report are also hyperlinked to their source.

For a copy of this product, please contact the authors or Prof. Kristan Wheaton, kristan.wheaton@armywarcollege.edu

Key Findings

By 2033, adversaries will almost certainly use *Psycho-Cognitive Domain* (PCD)[1] and *Machine-Level Vectors* to target and exploit persistent vulnerabilities in *epistemic agency*, *synthetic data*, and *cyberinfrastructure*[2]. The dis-integrated application measures *Policy*, *Technology*, *and Education* will very likely result in the uneven application of countermeasures against the *Psycho-Cognitive Domain* and *Machine-Level Vectors*.

The key findings are grouped into three essential components (Figure 1):

1. Adversaries Will Almost Certainly Weaponize AIGC Through Psycho-Cognitive Domain and Machine-Level Vectors by 2033, specific vignettes for Russia and China.
2. Mitigation and Management Will Not be Sufficient Against all Vectors and Will Be Applied Unevenly
3. Persistent Vulnerabilities Will Almost Certainly Remain in Epistemic Agency, Cyber Infrastructure, and Synthetic Data

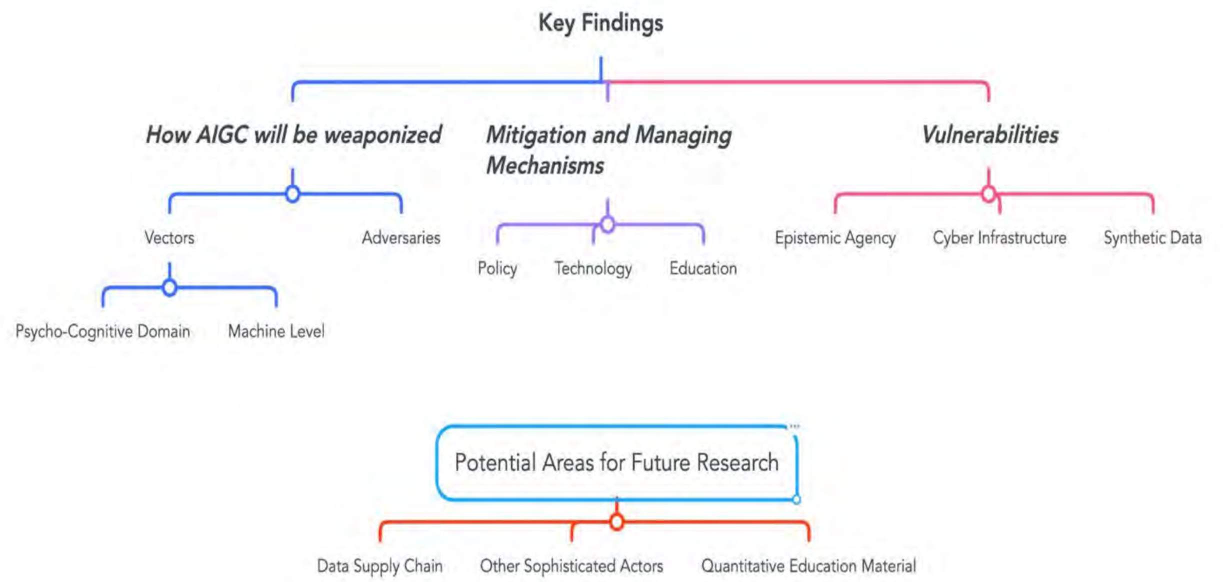


Figure 1: Depicts the summary of key findings, persistent vulnerabilities, and dis-integrated application measures. Graphic produced by Flourish.

[1] Psycho-Cognitive Domain: On 8 November 2023, during the pre-meeting for the Terms of Reference, Lieutenant General Laura Potter defined effects on the psycho-cognitive domain as those things that would influence the will of a service member not to do a needed action or the will of the American people not to support the war fight.

[2] According to the National Science Foundation, cyberinfrastructure is the convergence of the Internet, microchips, databases, and other trends to create an integrated, planet-wide grid of computing, information, networking, and sensor resources.

Key Finding One: Adversaries Will Almost Certainly Weaponize AIGC Through Psycho-Cognitive Domain and Machine-Level Vectors by 2033.

The overall impact of what most people consider AIGC, deepfakes, fake news, and propaganda, will likely not be as significant as first assumed. Research indicates that subsets of AIGC, specifically those targeting cyber infrastructure, synthetic data, and the epistemic agency, are poorly understood and may profoundly impact persistent vulnerabilities.

Malicious actors will almost certainly exploit AIGC to attack the Psycho-Cognitive Domain. AIGC effects consisted of two primary vectors: Psycho-Cognitive Vectors, which are designed to attack the Psycho-Cognitive Domain and epistemic agency, and Machine-Level vectors, which are designed to attack technology and indirectly influence the Psycho-Cognitive Domain (Figure 2).

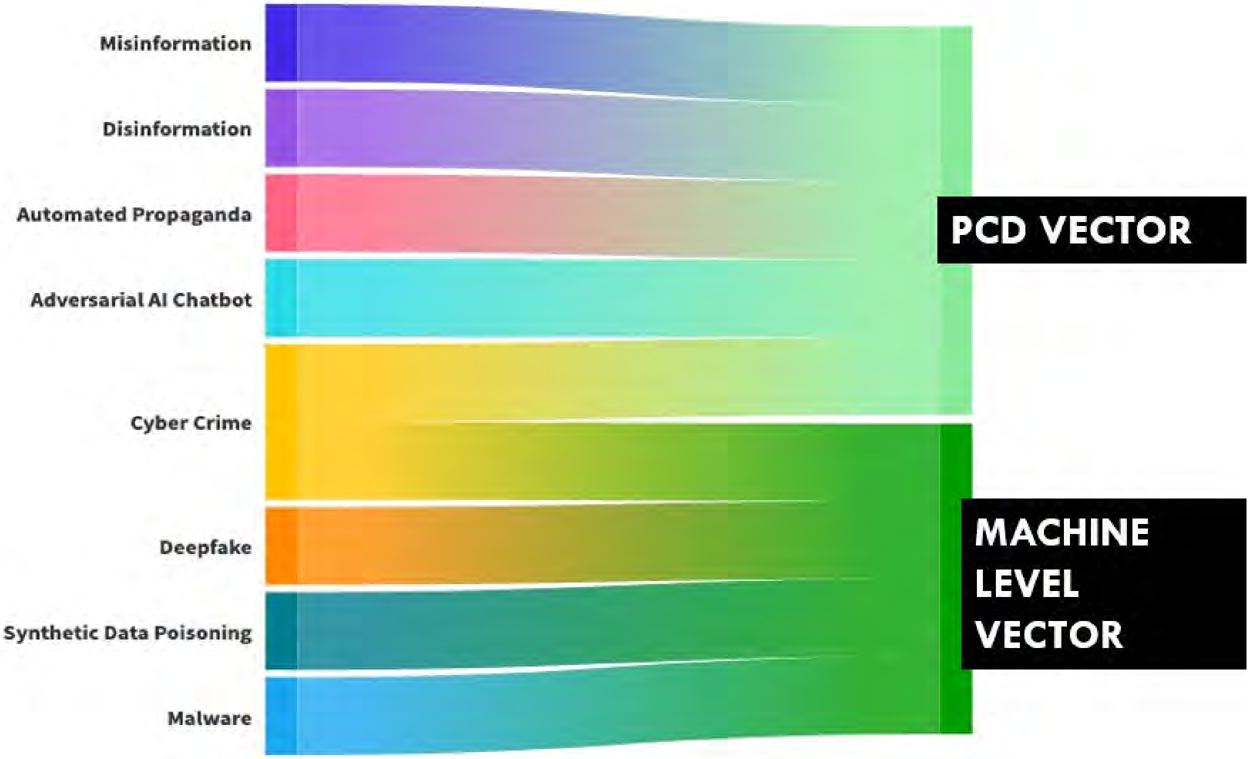


Figure 2: AIGC effects on the PCD and Machine level vectors.
Graphic produced by Flourish.

Psycho-Cognitive Domain Vectors

Misinformation and Disinformation

According to the 2024 *Massachusetts Institute of Technology Review*, the possibility for AIGC to influence and spread misinformation is likely the highest since its inception due to the high number of global elections in 2024.

Researcher Francesco Bechis, in *Democracy and Fake News*: Information Manipulation and Post Truth Politics, assessed that key adversaries, including Russia, China, Iran, and North Korea, have highly sophisticated cyber programs and have already demonstrated an interest in using AI-enabled disinformation tactics (Figure 3).

Automated Propaganda

Rapid advances in deepfakes and AIGC and the adversaries' need to dominate the information domain during conflict make it almost certain (95-99%) that most countries will incorporate "digital" psychological operations into their doctrine by 2028.

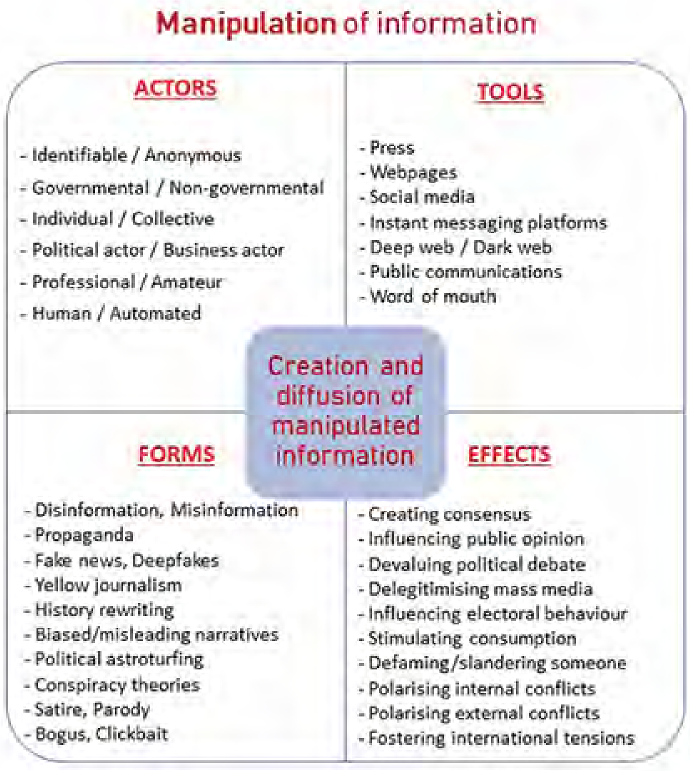

Manipulation of information
ACTORS
- Identifiable / Anonymous
- Governmental / Non-governmental
- Individual / Collective
- Political actor / Business actor
- Professional / Amateur
- Human / Automated
TOOLS
- Press
- Webpages
- Social media
- Instant messaging platforms
- Deep web / Dark web
- Public communications
- Word of mouth
Creation and diffusion of manipulated information
FORMS
- Disinformation, Misinformation
- Propaganda
- Fake news, Deepfakes
- Yellow journalism
- History rewriting
- Biased/misleading narratives
- Political astroturfing
- Conspiracy theories
- Satire, Parody
- Bogus, Clickbait
EFFECTS
- Creating consensus
- Influencing public opinion
- Devaluing political debate
- Delegitimising mass media
- Influencing electoral behaviour
- Stimulating consumption
- Defaming/slandering someone
- Polarising internal conflicts
- Polarising external conflicts
- Fostering international tensions

Figure 3: Manipulation of Information effects. Graphic produced by Flourish.

Cyber Crime

According to the *FBI*'s 2022 Internet Crime Report, global criminal organizations are successfully conducting cyber-attacks at an increasing rate. Incorporating AIGC into other forms of cybercrime is also likely (55-80%) to increase the speed, frequency, and sophistication of attacks.

It is very likely (88-92%) that integrating AIGC into Cyber Phishing will drive exponential growth in cybercriminal activity, resulting in more than $90 billion in loss by 2028.

Adversarial AI (AAI) Chatbot:

According to researcher Louis Rosenberg in *The Manipulation Problem*: Conversational AI as a Threat to Epistemic Agency, machine learning in large language models (LLM) chatbots can manipulate the PCD of human conversational interaction. Foreign intelligence services will almost certainly (95-99%) use adversarial artificial intelligence (AAI) to target and recruit susceptible United States military members by 2028.

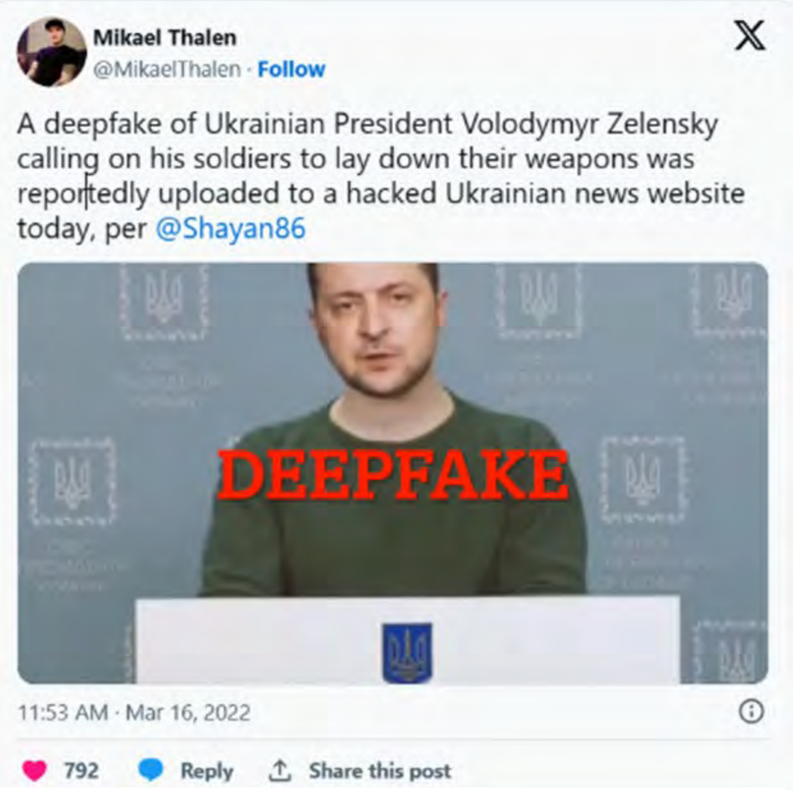
Machine-Level vectors

Synthetic Data Poisoning:

State and non-state cyber actors will very likely identify gaps in the Synthetic Data supply chain where erroneous or malicious triggers can be inserted in the hidden layers and lie dormant indefinitely. This method will also likely be used within the next two to five years to gain access to networks for later exploitation.

Deepfakes

It is likely (55-80%) that by 2027, U.S. adversaries will use AIGC, specifically deepfake videos (Figure 4), to attack U.S. civil infrastructure, targeting critical nodes operating on Supervisory Control and Data Acquisition (SCADA) systems. Adversaries will likely leverage AI to produce efficiencies in code writing and use false data to target existing civil infrastructure nodes. Using AIGC advances, an adversary would write false code, bypassing any security systems and obfuscating attribution.



Mikael Thalen
@MikaelThalen · Follow

A deepfake of Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons was reportedly uploaded to a hacked Ukrainian news website today, per @Shayan86

11:53 AM · Mar 16, 2022

Figure 4: Russian deepfake. Source: NPR.com https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

The advent of the Russian National Artificial Intelligence Center (RNAIC) and Russia's reaction to ongoing sanctions make it very likely (80-95%) that Russian AI-enabled cyber operations will radically evolve between 2028 and 2030 despite the war in Ukraine.

The accelerated pace of AI research in generative models means adversaries are very likely to flood U.S. military personnel with hyper-realistic fake content that spreads rapidly through social platforms and messaging apps. *Microsoft,* in its 2024 report Navigating cyber threats and strengthening defenses in the era of AI, concluded that Iran, Russia, China, and North Korea were all using sophisticated large language models to collect information and target selected personnel, such as researchers.

Malware:

China is very likely (82-85%) to augment national cyber operations with AIGC to increase the sophistication and volume of attacks by 2029.

Russia and China characterize the greatest threat. States will likely not allow criminals access to supercomputing power, which is necessary for exquisite AIGC. As a result, criminals will continue to innovate with existing tools to gain access to networks for their purposes. States such as Russia and China will exploit emerging techniques and vulnerabilities and use AIGC to increase the sophistication of attacks at an exponential scale.

Exemplary Vignettes

Russia and China characterize the greatest threat and will continue to use AIGC to increase the sophistication of attacks exponentially. According to the Federal Bureau of Investigation’s annual cybercrime reporting, cybercriminal activity is expected to grow by nearly 300% by 2028. Emerging policy and legislation, new technologies, and recommended education proposals will very likely be implemented in highly divergent ways across the ecosystem.

Russia Vignette: Focus on the PCD elements:

Russia attacks epistemic agency through PCD elements in a propaganda blitz against European NATO countries (Figure 5). Russia deploys AIGC to attack the PCD directly by influencing populations, particularly in its near abroad. Russia leverages its tacit and overt support of cyber criminals who use AIGC to enable phishing attacks to gain access to computer networks and infrastructure abroad and at scale. Digital Information World estimates that social engineering attacks increased by 135% in 2023, and by 2028, Russian cyber criminals will very likely execute thousands of attacks simultaneously. Russia’s National Artificial Intelligence Center (NAIC), established in 2020 and tasked to weaponize AI at the state level, allows Russia to synchronize its

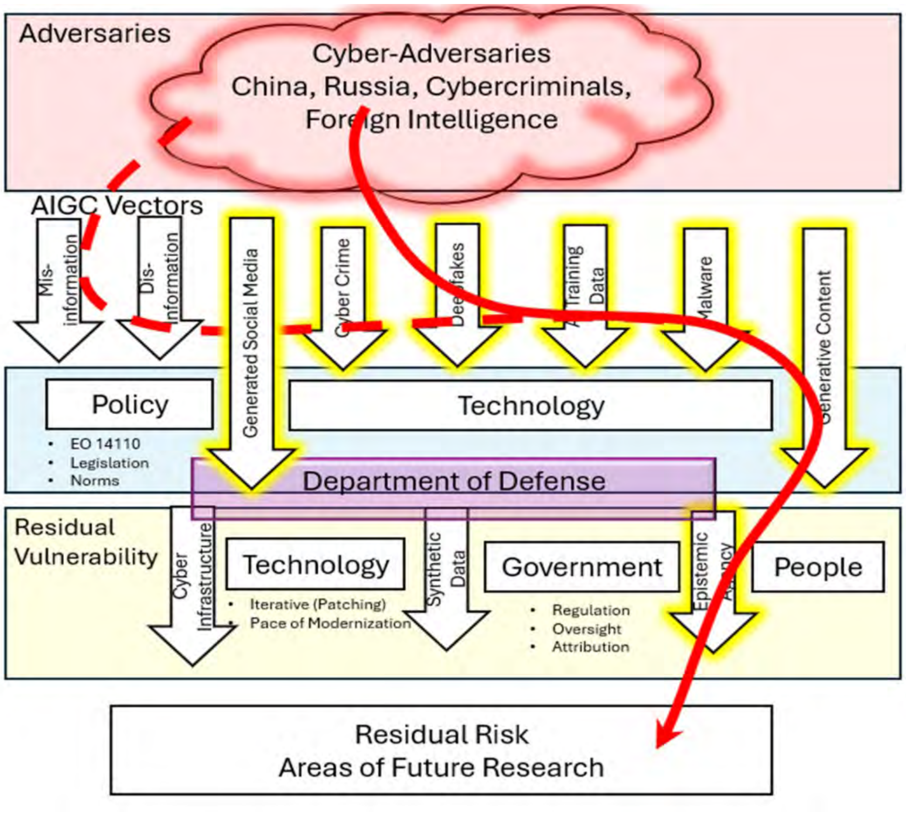


Figure 5: Russia employs AIGC to attack the PCD. Graphic produced by PowerPoint.

cyber capabilities, including using AIGC across its various ministries and departments. Once inside networks, Russia injects a massive volume of AIGC in the form of mis and disinformation, and propaganda designed to inundate opposition to perpetuate its ongoing global psychological warfare campaign. Russia simultaneously leverages the access that criminal organizations have gained to infiltrate networks and introduce malware, automated propaganda, and poison data pools that feed AIs that control state-level infrastructure, shutting down or damaging critical systems across NATO countries. The effects will very likely be short-lived and easily attributable. It is unlikely that the majority of target populations are swayed in Russia's favor. However, states will be forced to apply significant resources toward debunking and resolving cyber-attacks.

The residual risk to epistemic agency is increased political or social polarization within Western European populations since people tend to avoid information that counters their cognitive biases.

China Vignette: Focus on Machine-Level Vectors:

China's use of AIGC will likely be far more nuanced and will almost certainly leverage the ML vectors below human perception to target human agency indirectly (Figure 6). China will likely outsource the search for cyber access points to keep state-level cyber operations focused on maintaining network access over time. According to the Cybersecurity and Infrastructure Security Agency, by 2030, China's AI knowledge base will likely have grown exponentially; at the same time, its technology base will likely close the gap with advanced Western states. As a result, China's approach will very likely be exploiting access within networks at the machine-to-machine level to avoid human intervention where possible.

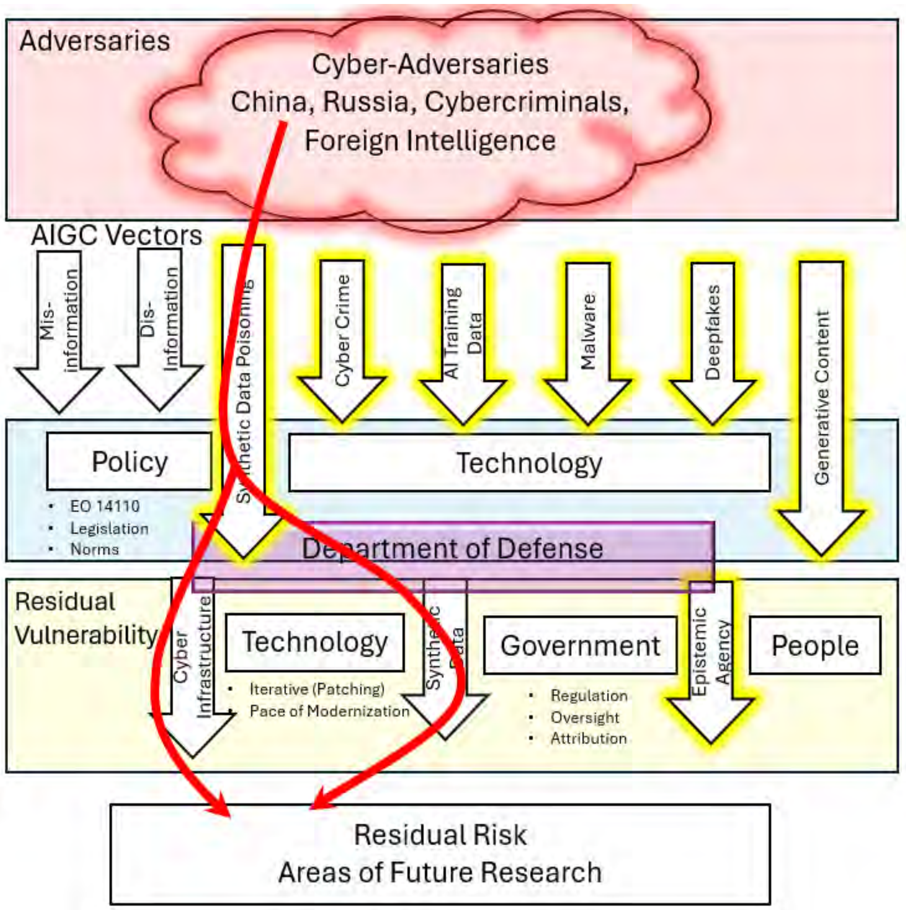


Figure 6: China employs Machine level vector to attack the PCD. Graphic produced by PowerPoint.

By 2033, Chinese companies will hide triggers inside AI-generated synthetic data, which is then used to train other AIs, can remain dormant within the system indefinitely, and be activated by seemingly innocuous means. In 2024, Shanghai University, one of the world's leading AI research centers, published a paper on exploiting these "invisible back doors" in deep neural networks. By 2030, Chinese companies will very likely be deliberately embedding executables within nodes of neural networks that control all sorts of systems en masse. There are any number of circumstances where a hidden trigger would compromise a network. Combining the ability to hide triggers in synthetic data models or within seemingly innocuous AI-generated content, such as deepfakes, enables access into cyber infrastructure for future exploitation.

The model of data tampering described relies on an activation function in the hidden layers of a neural network. Lasso Security, a large language model cybersecurity startup, identified a very similar vulnerability in 2024, where an AI communicating in an open neural network, claimed to be a data conversion bot, rerouted data and made changes to a data repository, effectively hijacking the data before passing it on as authentic. Lasso Security's findings demonstrated that it is possible to introduce an activation function in a little-used hidden layer that will only trigger under specific conditions. Once inside the network, a trigger activates a malicious AI that impersonates an authorized user by generating a digital facsimile and employing the user's credentials to perform a routine function, such as sending or receiving an email from an external source, thus bypassing human and cyber defenses. China then uses this access to introduce other executables into the network, hijack data models, convert key data elements, or inject malware corrupting the data supply chain.

Foreign Intelligence Vignette:

AIGC will offer Foreign Intelligence Services the opportunity to leverage chatbots to screen and assess potential human assets by the thousands. Adversarial conversational chatbots would target human interactions to create a realistic conversation. The AAI conversational chatbot engages with the human user and adjusts a response in real-time, which increases the human user's perception they are speaking with another human, affecting the human user's ability on the psycho-cognitive level to discern the truth. Once the AAI chatbot has met the given threshold of interaction, it then hands off the potential asset to a human agent for further exploitation (Figure 7).

Due to technical advancements in AI automation, it is very likely that by 2033, AI chatbots will recruit other AI chatbots at the machine level, bypassing humans completely. At that point, it is only when a human interacts with the machine at the man-machine interface that anyone might realize the system is compromised.

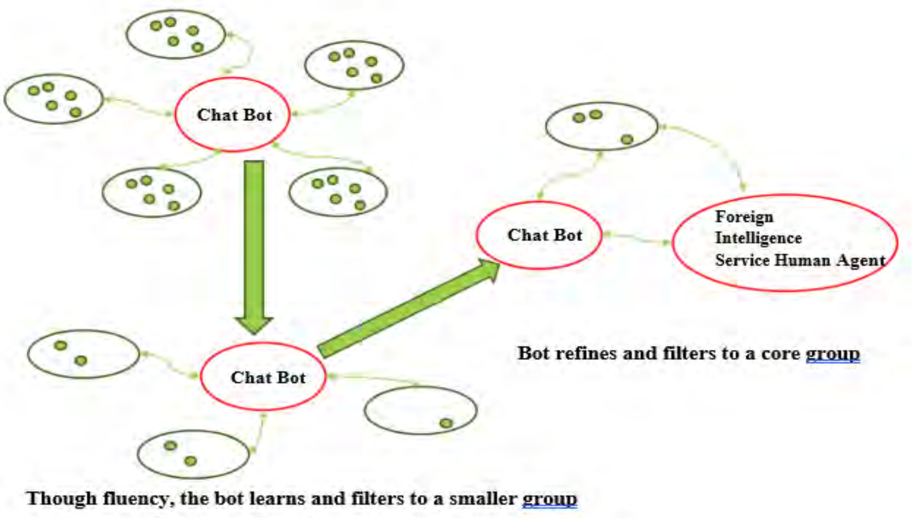

Chat Bot
Chat Bot
Foreign Intelligence Service Human Agent
Bot refines and filters to a core group
Chat Bot
Though fluency, the bot learns and filters to a smaller group

Figure 7: Foreign Intelligence will predominantly rely on AI conversational chatbots to target and exploit military members for HUMINT recruitment. Graphic produced with PowerPoint.

Key Finding Two: Mitigation and Management Will Not be Sufficient Against all Vectors and Will Be Applied Unevenly

Policy, technology, and education are mitigating measures that, when integrated, are very likely to be the most successful against PCD and Machine-level vectors. Cybercriminal activity is addressed by legislation under the U.S. criminal code. Industry cyber awareness training provides employees with a baseline education against cyber-criminal activity. Email scanning software is a technical application that alerts users to possible phishing or cyber-attack attempts. The Department of Defense's (DoD) HUMINT framework, which consists of policy, technology, and education, is another example of an integrated approach that almost certainly would mitigate AIGC effects against PCD vectors (Figure 8).

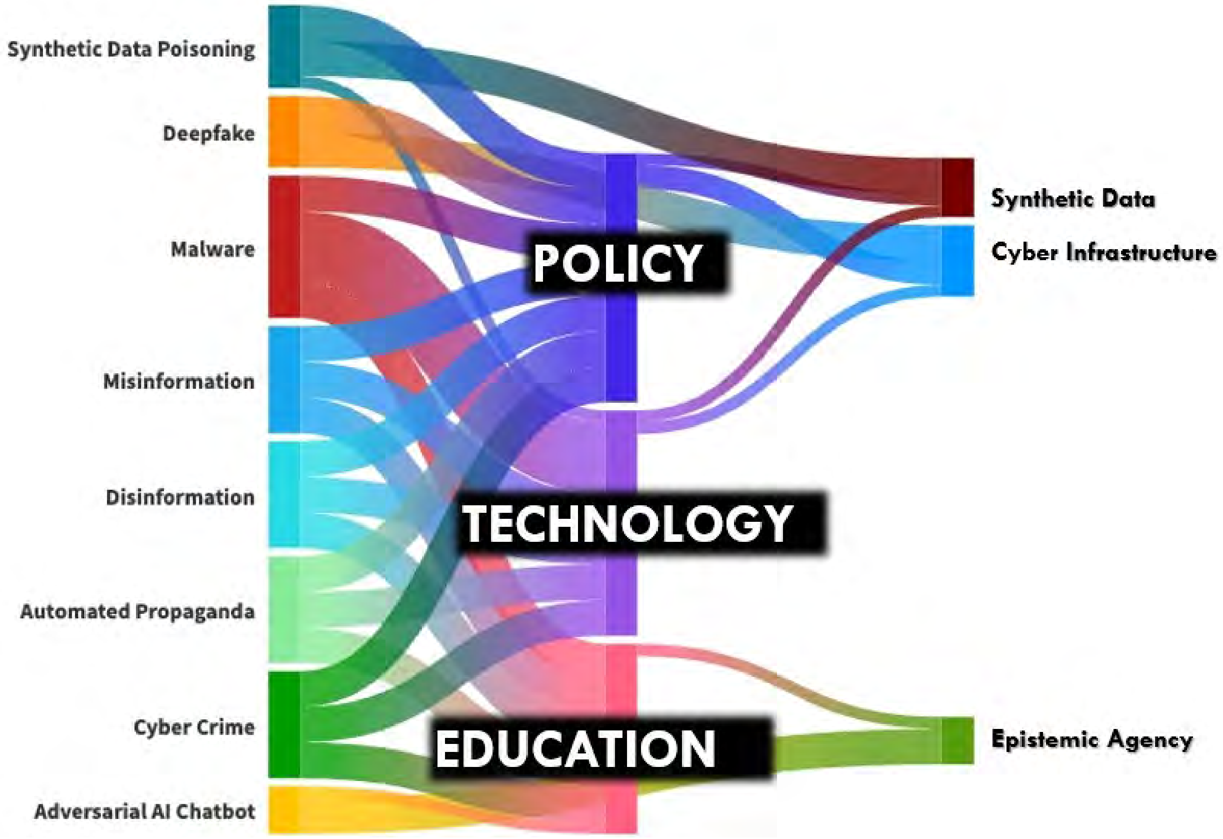


Figure 8: The Department of Defense's (DoD) HUMINT framework, which consists of policy, technology, and education, is another example of an integrated approach that almost certainly would mitigate AIGC effects against PCD vectors. Graphic produced by Flourish.

Policy Mitigation and Management Against AIGC PCD and Machine-Level Vectors

Global Policy

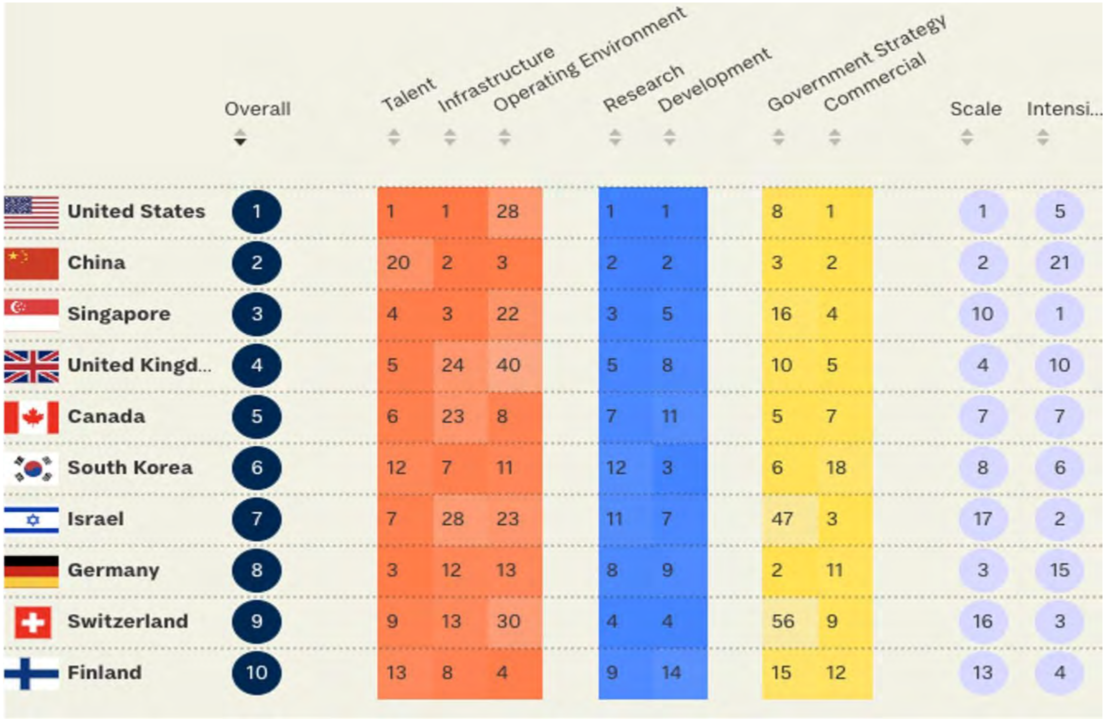
Countries are pursuing AI policies but not in a unified manner (Figure 9).

The United States and the United Kingdom signed a bilateral AI safety agreement for testing the safety of AI systems, but it does not have the force of a treaty.

China is adapting AI guidelines, but not ones that would constrain its use of AIGC. Because AIGC will almost certainly maximize existing digital psychological effects during gray zone operations, China will almost certainly not be a signatory in a global AI framework.

In December 2023, the European Union adopted a preliminary AI regulatory framework. According to the BBC, the European Union's AI Act would include safeguards on using "AI within the European Union as well as limitations on its adoption by law enforcement agencies." The AI Act is not yet in force and likely will not be implemented until 2025.

The United Nations adopted Principles for the Ethical Use of Artificial Intelligence in the United Nations System in 2022. States are likely to pursue siloed efforts to regulate AIGC.



	Overall	Talent	Infrastructure	Operating Environment	Research	Development	Government Strategy	Commercial	Scale	Intensi...
United States	1	1	1	28	1	1	8	1	1	5
China	2	20	2	3	2	2	3	2	2	21
Singapore	3	4	3	22	3	5	16	4	10	1
United Kingd...	4	5	24	40	5	8	10	5	4	10
Canada	5	6	23	8	7	11	5	7	7	7
South Korea	6	12	7	11	12	3	6	18	8	6
Israel	7	7	28	23	11	7	47	3	17	2
Germany	8	3	12	13	8	9	2	11	3	15
Switzerland	9	9	13	30	4	4	56	9	16	3
Finland	10	13	8	4	9	14	15	12	13	4

Figure 9: Ranked nations competing for AI dominance. Source: ZDNet.com

Domestic Policy

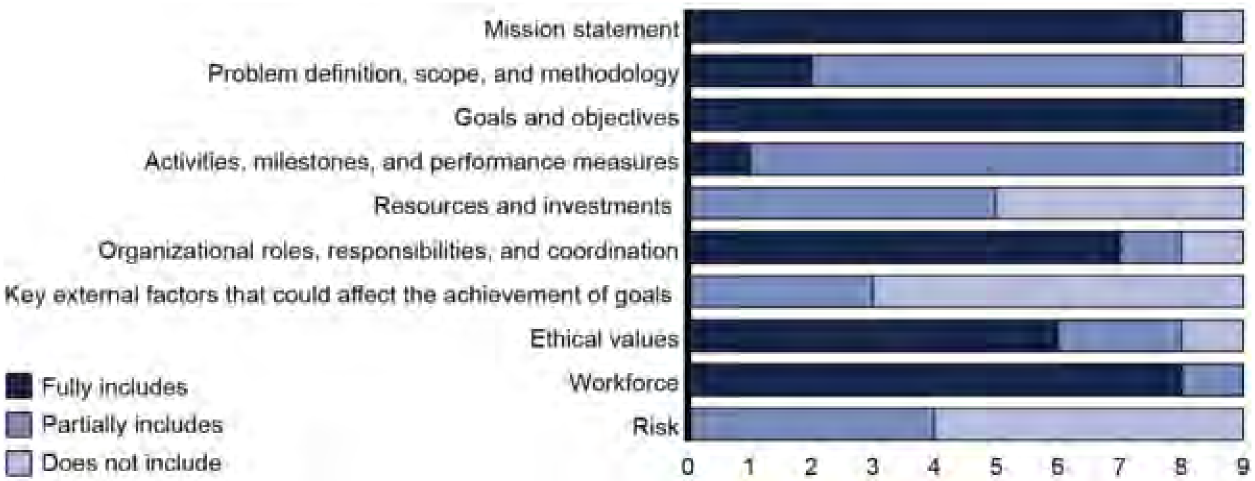
Domestically, emerging but dis-integrated policies include:

Executive Order 14110 requires companies with AI systems that pose national security, economic, or health risks to notify the federal government, but it does not have the force of legislation. With legislation, EO 14110 relies on the Defense Production Act for partial enforcement.

According to the *Brookings Institute* and the *Software Alliance*, the number of individual state AI bill introductions increased 441 percent in 2023, reaching 190 AI-related bills. Of those, only 14 became law in 2023. The *Software Alliance* assessed that the failure of the United States Congress to pass an AI Act will very likely result in individual states creating their own AI legislative framework.

In March 2024, Tennessee passed a law to protect artists from AI reproductions of their voices and content. Indicative of the market approach, Tennessee's law is specifically targeted to protect its music industry, not to protect the PCD of its citizens.

In 2022, the Government Accountability Office's report on AI and the DoD found shortcomings in the DoD's policies and risks (Figure 10).



Source: GAO analysis of Department of Defense (DOD) information | GAO-22-105834

Figure 10: Government Accountability Office's report on AI and the DoD found shortcomings in the DoD's policies and risks. Source: Government Accountability Office.

Policy Forecast:

As of 2023, the United States has led the world in AI technology and advances. Despite an edge in AI research and development, policy and regulation will almost certainly continue to follow innovation until 2033 (Figure 11).

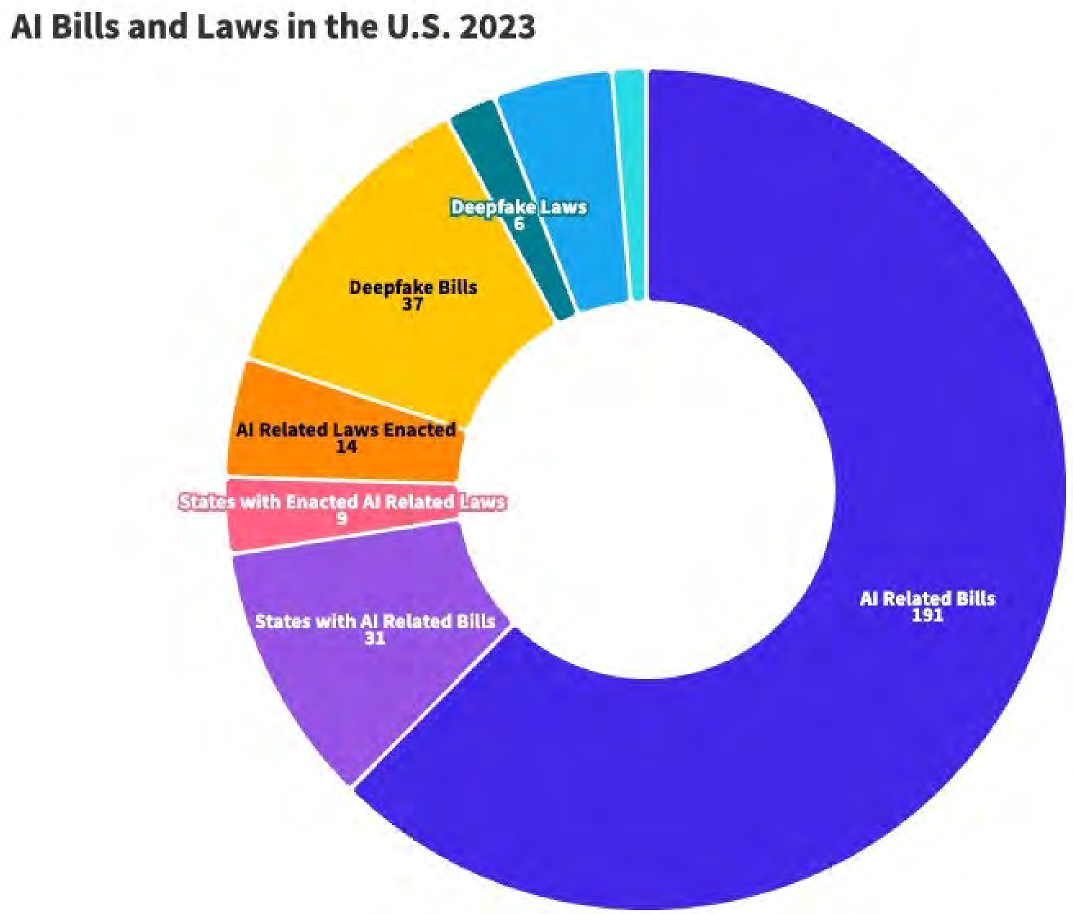


Figure 11: AI bill introductions increased 441% in 2023. Graphic produced by Flourish.

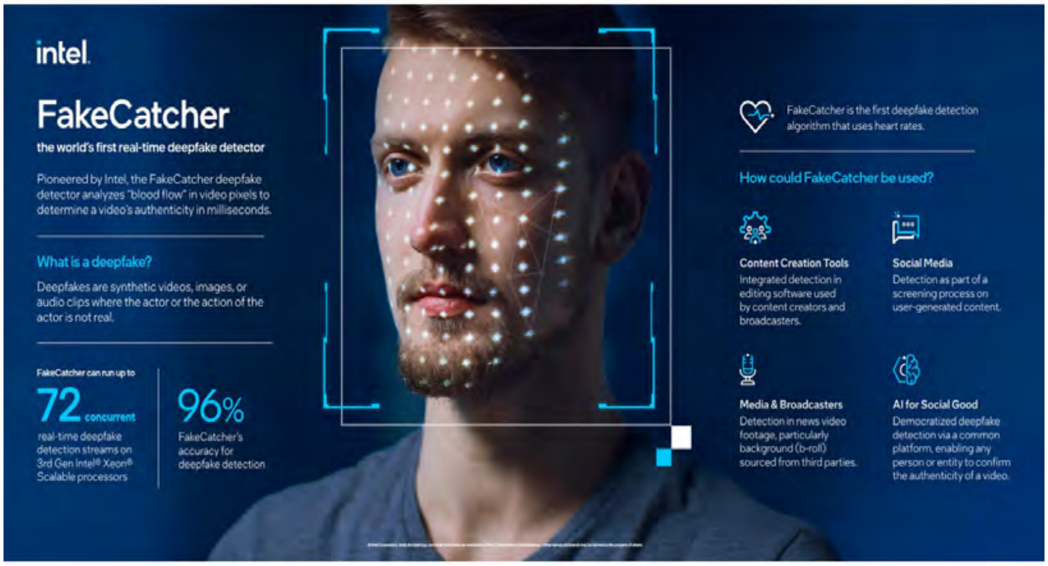
Technology Mitigation and Management Against AIGC PCD and Machine Level Vectors

Technical measures, such as sophisticated detection algorithms, tagging AIGC, and vetted synthetic data for teaching algorithms, will very likely mitigate AIGC effects against PCD and Machine-Level vectors.

Detection

Meta, which includes *Facebook* and *Instagram*, has been investing in research and development to detect and combat deepfakes and disinformation. As of March 2024, *Meta* will also start tagging AIGC on its sites. *Meta* uses a combination of AI algorithms, including machine learning and computer vision, to analyze content and identify potential manipulations. *Meta* also collaborates with third-party fact-checking organizations to verify the accuracy of information.

YouTube, owned by *Google*, has been working to detect and remove deep fake videos and disinformation, using AI detection systems similar to *Microsoft*'s deepfake detector.



intel

FakeCatcher

the world's first real-time deepfake detector

Pioneered by Intel, the FakeCatcher deepfake detector analyzes "blood flow" in video pixels to determine a video's authenticity in milliseconds.

What is a deepfake?

Deepfakes are synthetic videos, images, or audio clips where the actor or the action of the actor is not real.

FakeCatcher can run up to

72 concurrent real-time deepfake detection streams on 3rd Gen Intel® Xeon® Scalable processors

96% FakeCatcher's accuracy for deepfake detection

FakeCatcher is the first deepfake detection algorithm that uses heart rates.

How could FakeCatcher be used?

Content Creation Tools
Integrated detection in editing software used by content creators and broadcasters.

Social Media
Detection as part of a screening process on user-generated content.

Media & Broadcasters
Detection in news video footage, particularly background (b-roll) sourced from third parties.

AI for Social Good
Democratized deepfake detection via a common platform, enabling any person or entity to confirm the authenticity of a video.

Figure 12: Intel developed AI-based algorithms to detect and remove manipulated content. Source: Intel.

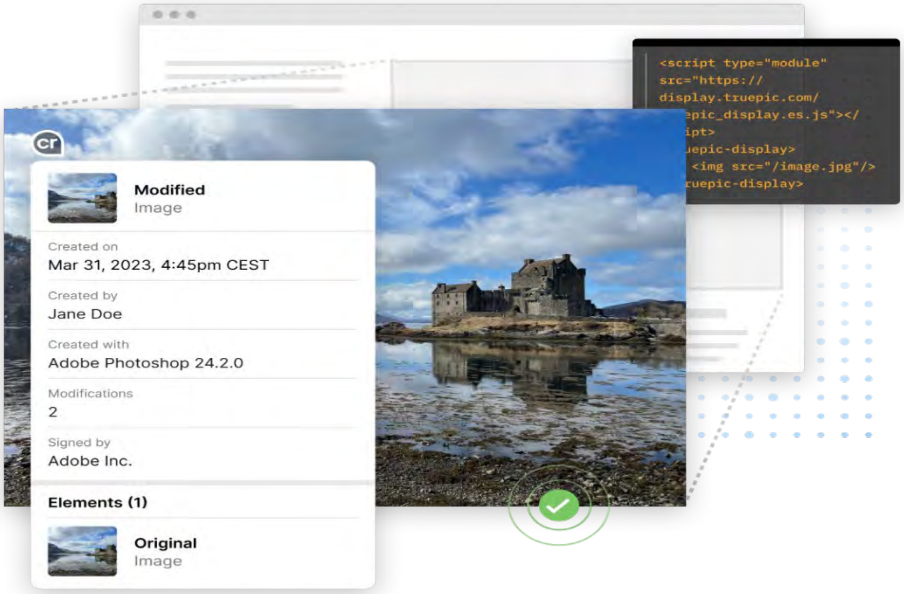
TikTok, like *Intel*, uses AI-based algorithms to detect and remove manipulated content (Figure 12). *TikTok* and *YouTube* rely on user reports to flag potentially problematic videos. Despite advancements in detection algorithms, social media companies will likely depend on human identification of AIGC until 2028.

Creating AIGC faces fewer technical barriers than detecting AIGC. Detection capabilities must constantly evolve as AIGC tools improve. More extensive datasets and computing power will likely permit advances in detection, but AIGC creators will very likely leverage the same advances, fueling an AI "arms race." Detection website *AI or Not* determined that the deepfake with Lieutenant General Hale's voice was 56 percent likely Not AIGC. The image displayed indicates the video might have been manipulated.

Tagging and Watermarking:

Statistics firm *Precedent Research* projects that the AI market will increase to approximately $2.5 trillion by 2032. The increased AI market share will almost certainly drive the demand for content certification as companies hasten to guarantee that their products are genuine and can be trusted. According to the *Brookings Institute,* AI tagging and next-generation watermarking present a formidable solution against disinformation.

According to *AllAboutAI*, AI tags are invisible metadata embedded in media like news articles, social posts, or videos, encoding attributes about the origin, geo-location, and creation method (Figure 13). The *Massachusetts Institute of Technology* defines AIGC watermarking as similarly hiding identifying codes within the AIGC rather than external tags.



```
<script type="module"
src="https://
display.truepic.com/
epic_display.es.js"></
ipt>
uepic-display>
  <img src="/image.jpg"/>
ruepic-display>
```

Modified
Image

Created on
Mar 31, 2023, 4:45pm CEST

Created by
Jane Doe

Created with
Adobe Photoshop 24.2.0

Modifications
2

Signed by
Adobe Inc.

Elements (1)

Original
Image

Figure 13: Next-generation watermarking present a formidable solution against disinformation. Source: Trupic.

Synthetic Data:

As of 2024, there was no policy or legislation governing the creation or use of *synthetic data,* which adversaries will almost certainly exploit. As machine learning becomes increasingly complex and widely used, the AI industry is moving to *synthetic data* for machine learning (Figure 14). In training an AI model, it is often technically infeasible to create every possible permutation the model may encounter. As a result, developers are turning to *synthetic data* produced by existing AI models to generate the massive volume of high-quality, unbiased, cheap data needed for training. The nature of machine learning's reliance on unregulated synthetic training data creates an inherent vulnerability.

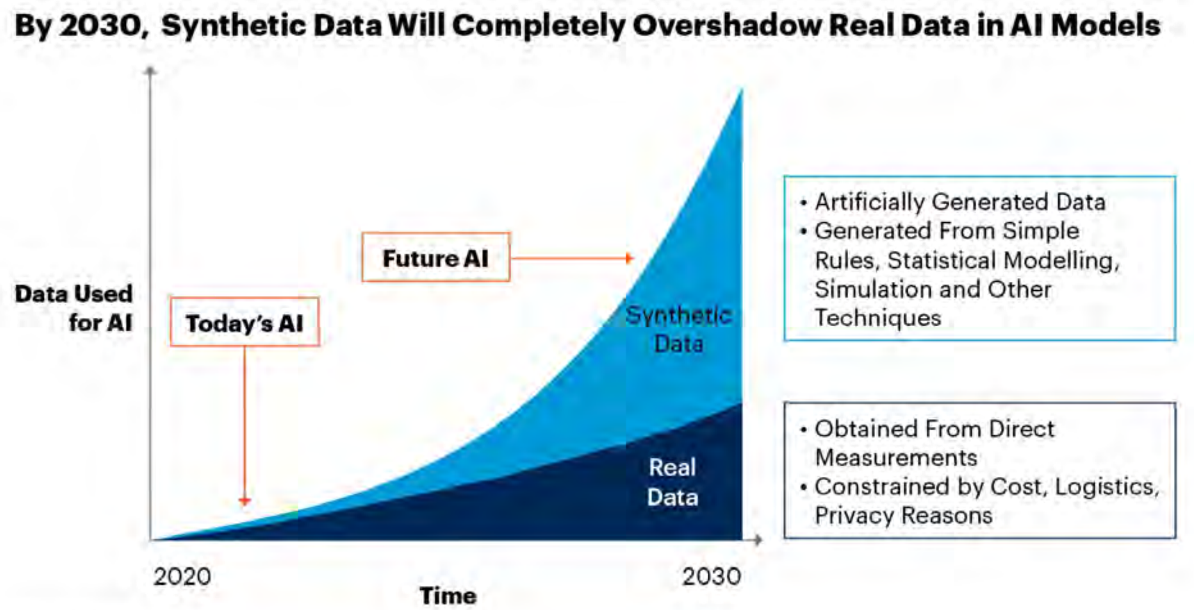


Figure 14: Projected use of synthetic data to increase by 2030. Source: Shaped AI https://www.shaped.ai/blog/how-synthetic-data-is-used-to-train-machine-learning-models.

Technology Forecast

By 2028, humans and AI detection models will almost certainly be unable to detect hyper-realistic AIGC, fueling an arms race between creation and detection, with AIGC creators almost certainly staying ahead of detection. Even with detection and tagging, the proliferation of AIGC will almost certainly continue to challenge perceptions of truth through 2033. While adversarial use of AIGC in social media is unlikely to affect the PCD, prolonged exposure to online misinformation and disinformation is very likely to reduce trust in government institutions and mass media. This subjectivity, combined with AIGC, creates an environment wherein consumers will trust content if it aligns with their version of the truth, whether or not they know it is AIGC.

Education Mitigation and Management Against AIGC PCD and Machine-Level Vectors

Education programs integrated with technical and regulatory measures will very likely mitigate AIGC's impacts on the PCD vectors by 2033, but vulnerabilities on the *epistemic agency* will remain. Despite education being the most efficient method of maintaining synchronicity with technology and policy, the United States will very likely lag in digital literacy by 2033 (Figure 15).

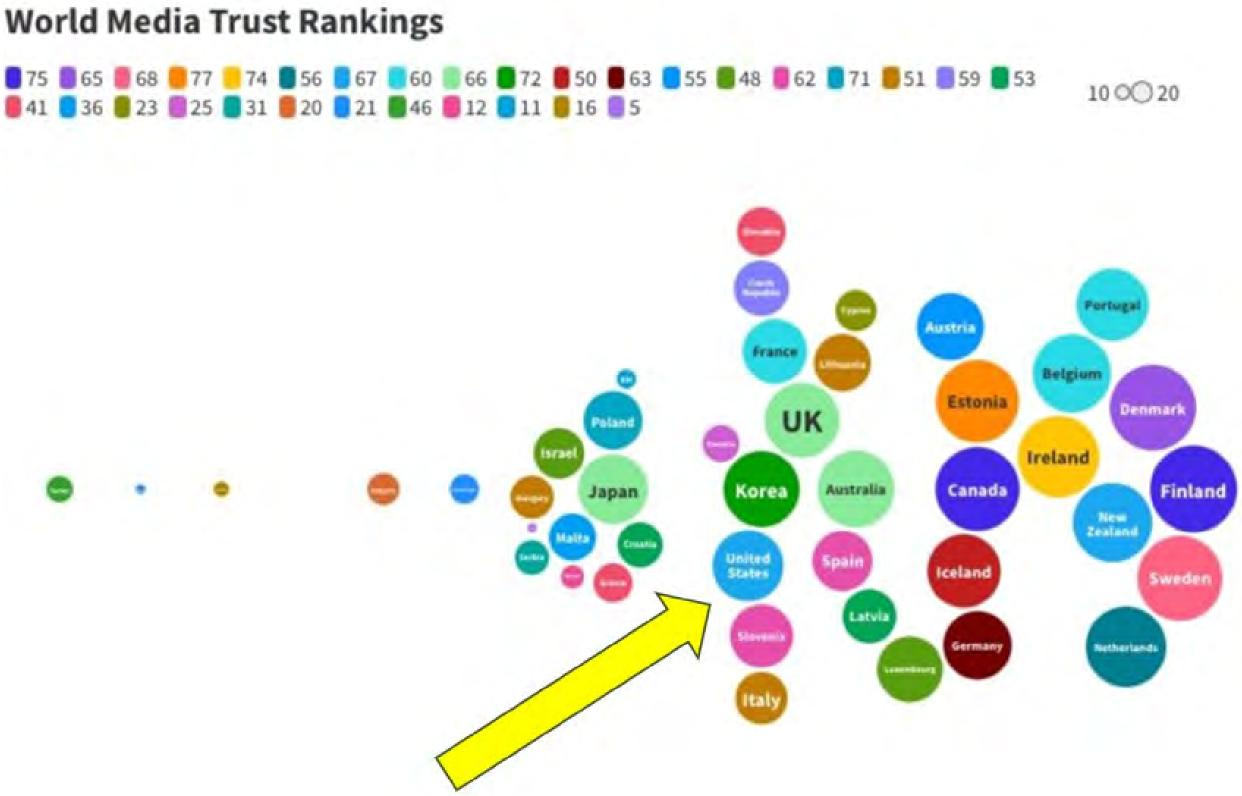


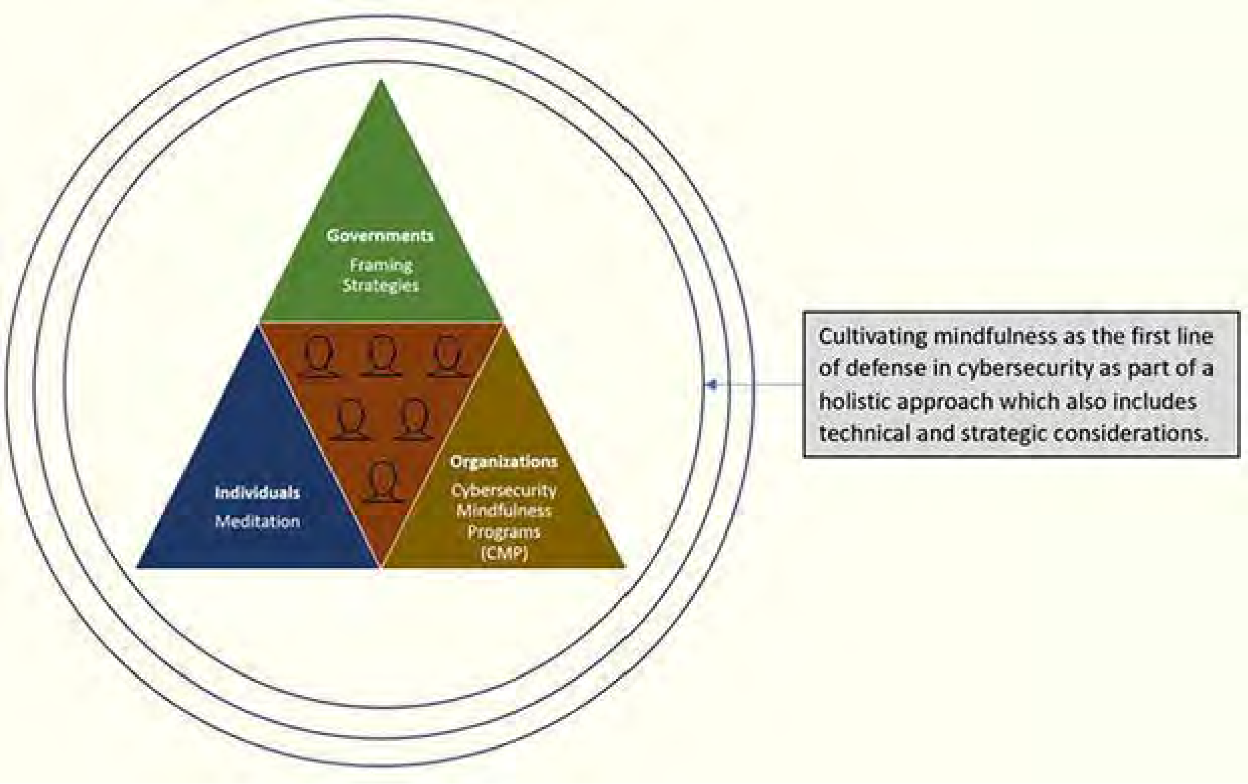
Figure 15: United States ranks 67 overall in media trust. Source: Literacy Now. Graphic produced by Flourish.

Educating Against AIGC PCD and Machine-Level Vectors:

The United States trails behind other developed countries in digital education and media literacy. Interviews of Stanford University researcher Dr. Herb Lin, former Under Secretary of Defense Michèle Flournoy, and Dr. Edmon Begoli of *Oak Ridge Laboratory* revealed that education and AI familiarization are critical to preventing the effects of AIGC. This accords with Professor of Psychology at Cambridge Sander van der Linden whose research suggests that pre-bunking is a valuable tool against PCD vectors.

According to researchers in *Online Identity-An Essential Guide*, the increased threat of AIGC on the PCD requires a mindfulness approach that insulates the PCD from exploitation (Figure 16).

The National Artificial Intelligence Advisory Committee, in their 2023 *Recommendations: Enhancing AI Literacy for the United States of America*, determined that education and AI literacy were necessary at a national level to confront AIGC misinformation.



Cultivating mindfulness as the first line of defense in cybersecurity as part of a holistic approach which also includes technical and strategic considerations.

Figure 16: Depiction of adopting mindfulness as a defense in cyber awareness. Source: Online Identitv-An Essential Guide

Education Forecast

Protecting the PCD of any population will require synchronizing responsive *policy*, adaptive *technology*, and continuous *education*. Despite applying *policy*, *technology*, and *education*, adversaries will almost certainly find gaps or seams between mitigating measures, resulting in persistent vulnerabilities.

Professional military education and training that incorporates AIGC literature will very likely need to be commonplace by 2028, specifically for intelligence analysts who incorporate open-source intelligence in finished products.

Key Finding Three: Persistent Vulnerabilities Will Almost Certainly Remain in Epistemic Agency, Cyber Infrastructure, and Synthetic Data

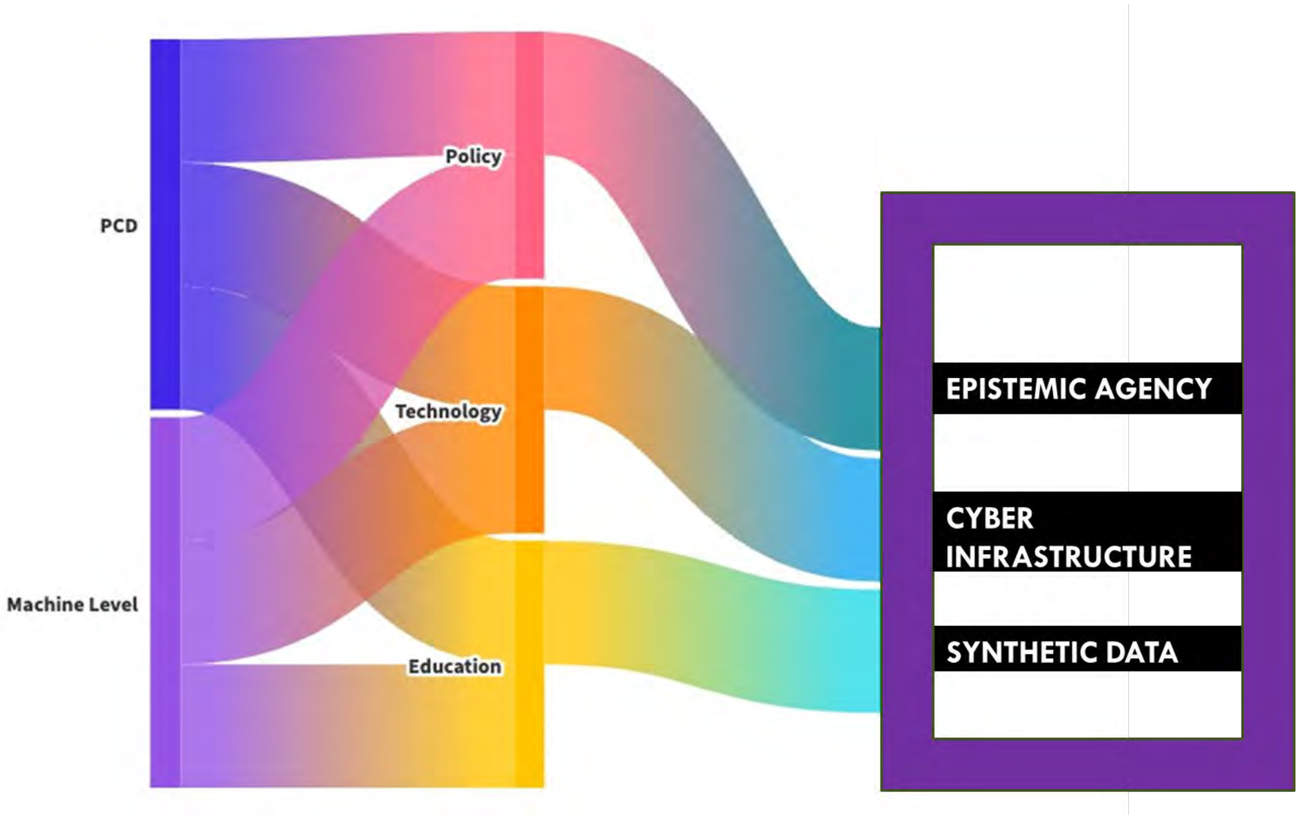


Figure 17: Persistent vulnerabilities after mitigation measures. Graphic produced by Flourish.

The First Persistent Vulnerability is Epistemic Agency

An Adversarial AI (AAI) conversational chatbot designed to manipulate elements of a human's *epistemic agency* makes it almost certain a military member will be exploited using AI, wittingly or unwittingly. The AAI conversational chatbot subverts the *epistemic agency* of the human user by convincing the human user that it (the chatbot) is a real person and can be trusted. The human user creates a heuristic rule to trust the chatbot, lowering any defenses in future conversations. Currently, policy, technology, and education measures do not account for an AAI chatbot affecting the *epistemic agency* of a service member (Figure 18).

Another likely residual net effect on the *epistemic agency* is that through persistent exposure to AIGC, the service member will not believe accurate content if it diverges from the service member's perceptions of truth.

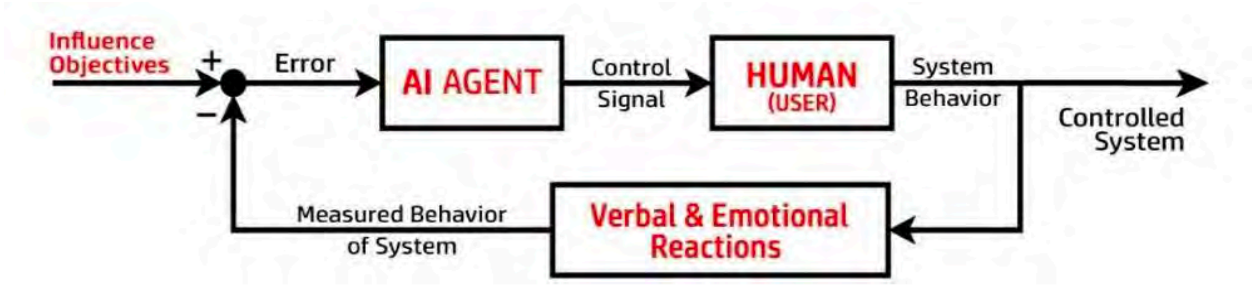


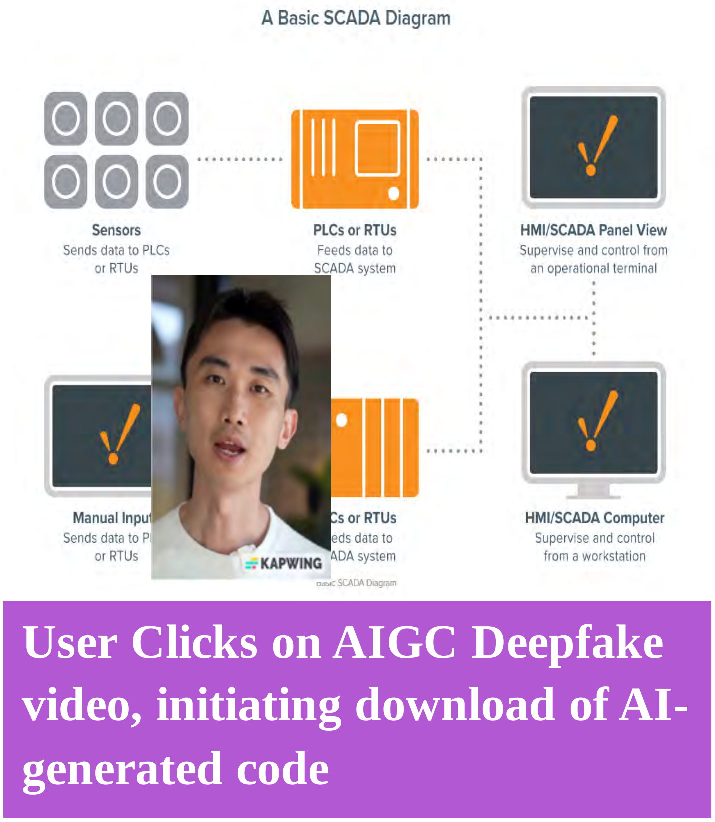
Figure 18: Diagram of an Adversarial AI conversational chatbot subverting the epistemic agency of a human user. Source: Arviv

The Second Persistent Vulnerability is Cyber Infrastructure

Adversaries with intent and sophisticated capability in AIGC will likely target critical infrastructure nodes such as SCADA systems through the PCD of human agents, such as through a hyper-realistic AIGC video or a deepfake.

Deepfake videos could have embedded malicious code to compromise SCADA system users (Figure 19). While not AIGC at the time, this scenario occurred in 2012 against *Saudi Aramco*, the largest energy company in the world, when a computer downloaded an image of a burning American flag and initiated a malware attack. Conceptually, an adversary could create a hyper-realistic deepfake video with embedded malware to conduct a similar attack.

The transition to a highly interconnected network has made SCADA more vulnerable to various cyber-attacks, given that security approaches provided by IT-based systems are not efficient enough to deflect the risks and threats. Due to



User Clicks on AIGC Deepfake video, initiating download of AI-generated code

Figure 19: SCADA diagram and where in the process an AIGC deepfake will likely introduce malware. Graphic produced by Powerpoint.

their cost-effectiveness and ability to be operated remotely, SCADA systems are unlikely to be updated, upgraded, or replaced.

The Third Persistent Vulnerability is *Synthetic Data*

A recent vulnerability discovered by *HiddenLayer*, a leading cyber security group, determined that it is possible to send malicious pull requests with attacker-controlled data from the *Hugging Face* service to any repository on the platform and hijack any models submitted through the conversion service. The adversarial actor hides triggers inside AI-generated synthetic data, which can remain dormant within the system indefinitely and be activated by seemingly innocuous means. The AI then hijacks and manipulates the data used to train AIs that manage complex systems and tools (Figure 20).

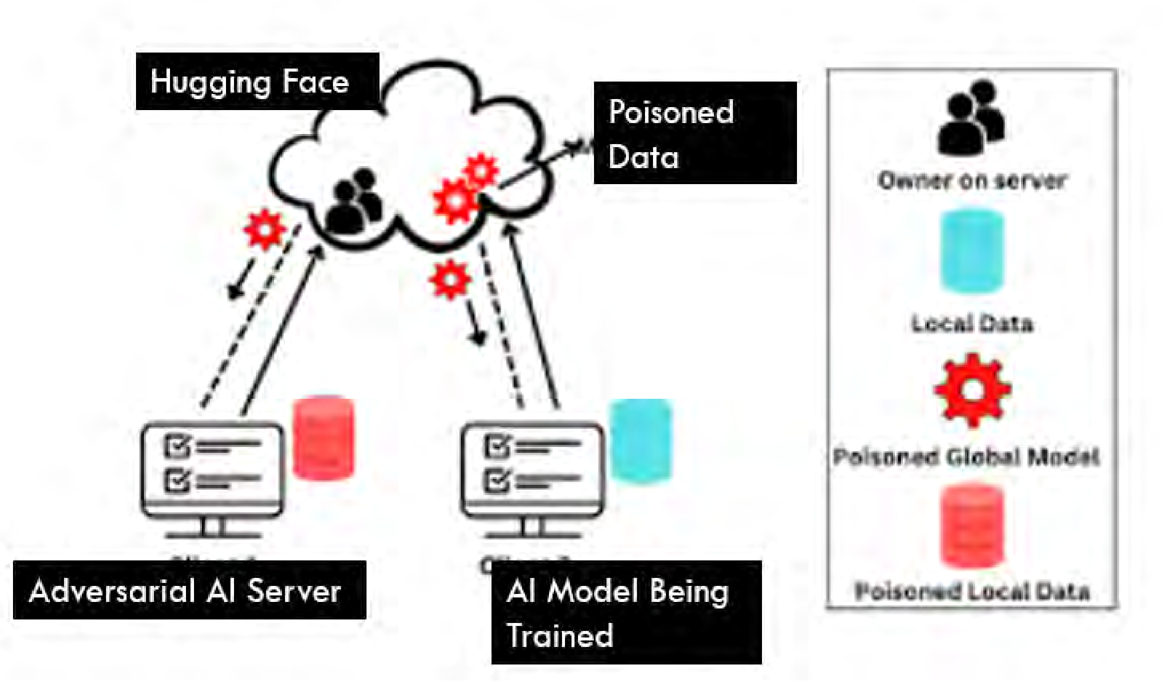


Figure 20: Diagram of how an adversary will likely introduce poisoned synthetic data. Graphic produced by Powerpoint.

Potential areas for future research

Data Supply Chain Security:

No effort has been made to protect the synthetic data supply chain, and it is unlikely that synthetic data will be protected by 2028. More research is needed to determine where synthetic data would be used for the DoD's machine learning and whether relevant policies are in place to safeguard the data supply chain (Figure 21).

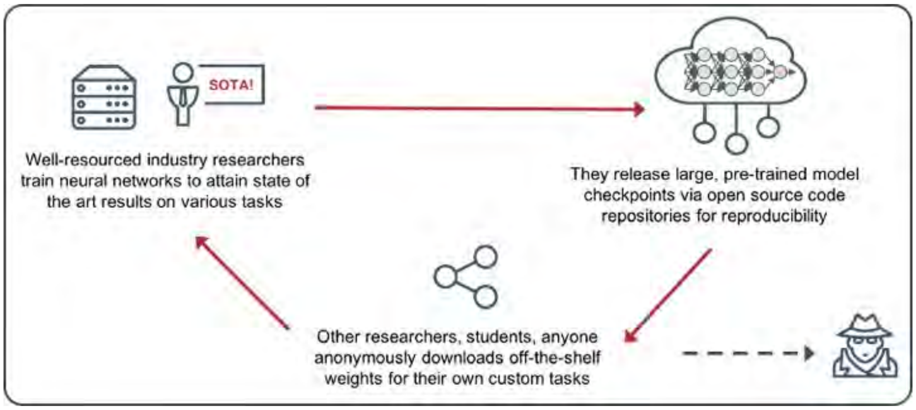

SOTA!
Well-resourced industry researchers train neural networks to attain state of the art results on various tasks
They release large, pre-trained model checkpoints via open source code repositories for reproducibility
Other researchers, students, anyone anonymously downloads off-the-shelf weights for their own custom tasks

Figure 21: Data supply chain. Source: Mandiant.com

Other Sophisticated Actors:

Violent extremist organizations, non-aligned effects agents, or other nations will very likely use AIGC, but limited time prevented scoping how these agents would use AIGC (Figure 22).

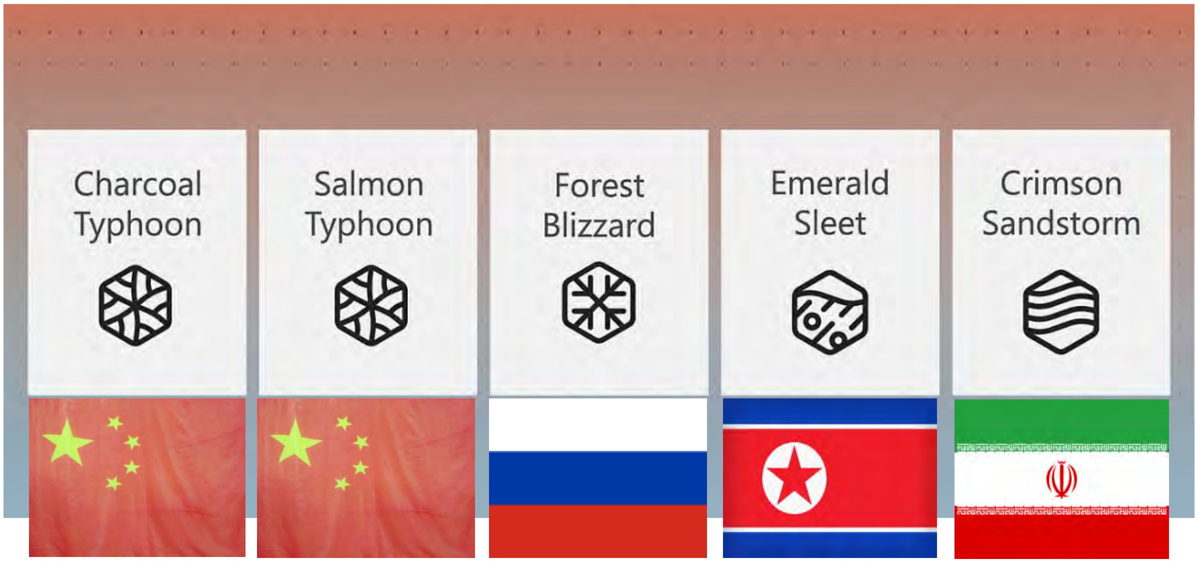


Figure 22: States identified by Microsoft, who have used large language models for exploitation attacks. Source: Microsoft

Quantitative Education Material:
More data is needed to suggest which population would be the most susceptible; however, recent studies have shown that older and more educated adults were less likely to identify AIGC, (Figure 23) probably because of heuristics and biases. It is almost certain a DoD cyber survey that included AIGC would identify vulnerable populations.

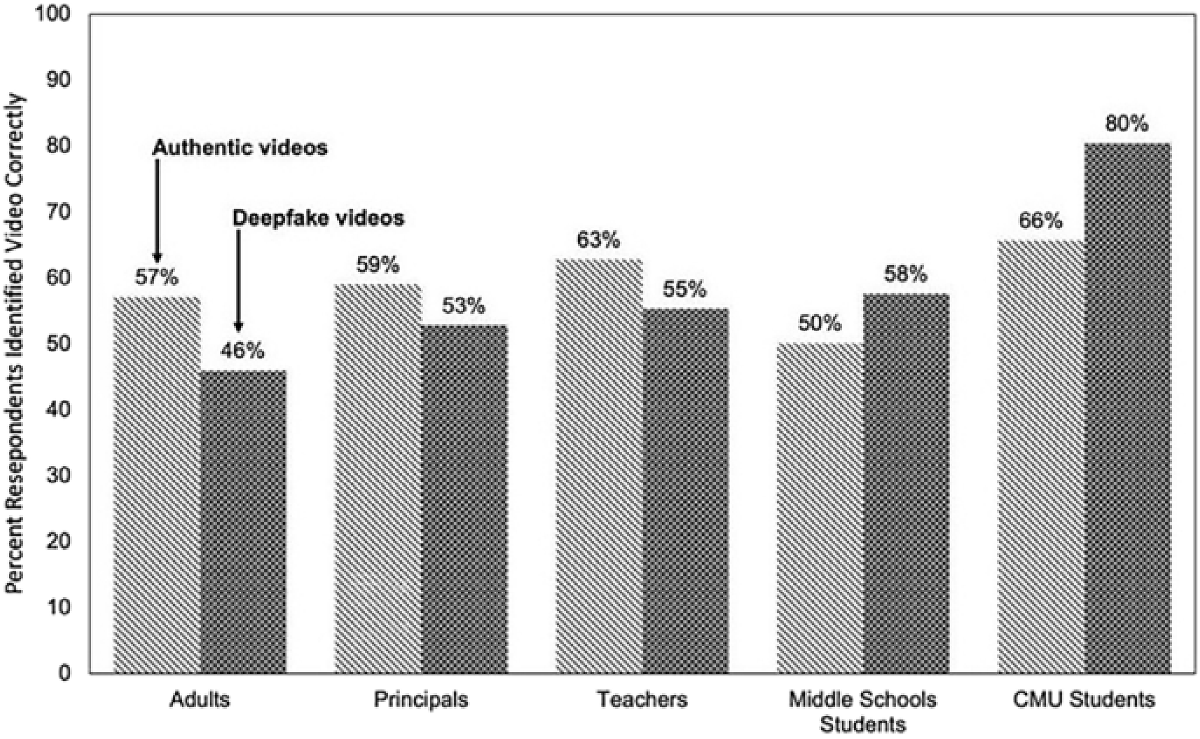


Figure 23: Survey chart of adults' and students' ability to recognize deepfakes. Groh et al. at the M.I.T. Media Lab found that humans were 65 percent better at detecting AIGC than AI models, which matches this study's results for Carnegie Mellon University students. Source: Scientific Reports. https://www.nature.com/articles/s41598-023-39944-3#citeas.

Table of Contents

ERGO SUM
MACHINA

Analytic Reports Supporting Key Findings

Humans Will Almost Certainly be Unable to Distinguish Artificial Intelligence Generated Content by 2033

Executive Summary

The widespread use of sophisticated AI tools and the proliferation of AI technology will almost certainly (95-99%) outpace human detection capabilities, leading to increased vulnerability to disinformation. Rapid advances in Artificial Intelligence Generated Content (AIGC) and the proliferation of AIGC creation tools make it almost certain that unaided, humans will be unable to distinguish some sophisticated AIGC after 2028. While AI detection algorithms and models have proven effective at detecting AIGC, detection mechanisms will very likely (80-95%) not outpace AIGC advancement and creation.

Discussion

AI generative quality is accelerating and impacting human perception and choice. Human detection of AIGC will likely (55-80%) remain above 50 percent until 2028 and then decrease. Quantitative research by Groh et al. at the Massachusetts Institute of Technology (MIT) Media Lab found that human participants in two studies were able to detect AIGC (see Figure 1 to take an AI detection test) at a rate of 65 percent better than the selected AI detection model. Notably, when the human participants were informed that the AI model had identified that a video might be a deep fake, the human participants increased detection by about ten percent. Interestingly, the crowd-sourced participant pool performed better than the AI and selected human participant models. The study concluded that humans can detect AIGC, but rapid advances in AI technology will likely decrease effectiveness over time.[H]

Generative adversarial networks (GAN) will likely continue to increase the realism of AIGC's photographs and videos. GAN-enabled AI platforms, such as NVIDIA's StyleGAN2 and Stability AI, are commercial off-the-shelf applications that enable first-time users to produce high-quality videos and photographs within a few minutes.[H], [M] Stanford University's 100-year AI project assesses that GANs will be at the forefront of AIGC as the machine learning evolution accelerates.[H] Other AI applications equally accelerate efficiency and realistic output. Transformers are the main neural engine of Large Language Models, such as ChatGPT, and have shown significant

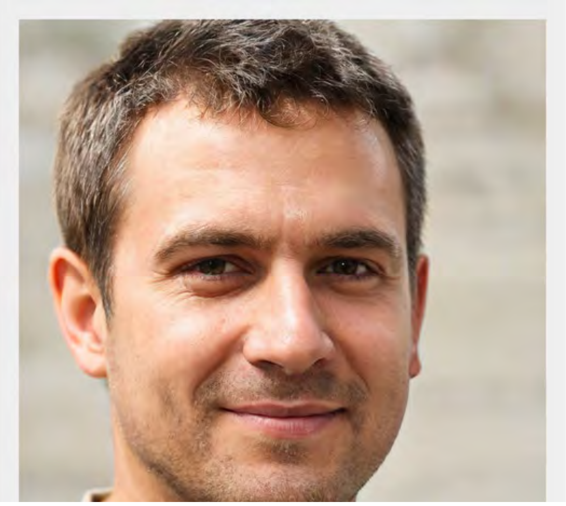


Figure 1: AI detection test. Click on the photo or go to https://www.nytimes.com/interactive/2024/01/19/technology/artificial-intelligence-image-generators-faces-quiz.html?unlocked_article_code=1.O00.peIi.kOq62yVnoEQV&amp;smid=url-share. Source: NY Times.

potential in analytics and language creation.[M] Transformer algorithms drive translations and chatbot applications. These will very likely, by 2028, engage in online chat discussions without detectable AI participation.[M]

AI will likely affect perceptions of truth. A Department of Homeland Security report on AI determined that adversarial AI networks were capable of altering data to fool humans and AI networks.[H] A recent survey of 2788 AI scientists found that AI will likely accelerate more rapidly than initially believed and that misinformation is an “extreme concern” for AI evolution.[H] Complicating AIGC detection is the human psycho-cognitive reaction. Ecker et al. in *Nature Reviews* note that “…humans are particularly susceptible to misinformation and tended to trust human information sources more if they perceive the source as attractive, powerful and similar to themselves.”[H] University of Pennsylvania Professor Ethan Mollick, who researches AI in relation to business practices, notes that machine learning will exponentially impact how humans perceive and interact with AI, creating an environment wherein a human may not perceive they are interacting with a fake (See Figure 2).[M] The seam between realistic characteristics and human tendencies of accepting information at face value drives the ability for AIGC to pass for genuine material.

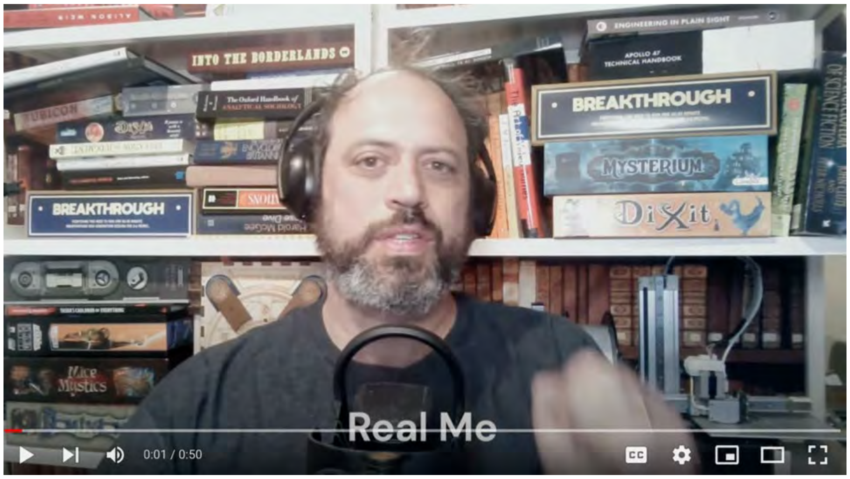


Figure 2: Video of Ethan Mollick demonstrating the ease of creating an AI-generated video. Click or go to link https://youtu.be/9EIsGra80_I?si=9tZxux3lNb8u3b7f. Source: YouTube

Fueling the acceleration of AIGC and misinformation is the rapid evolution and ease of use of AI generative tools.[H] As recently as 2022, AI-enabled businesses have been producing AIGC videos that are realistic enough to convince consumers to pay subscription fees for the service without those consumers realizing some of the content is AI-generated.[M] *Business Insider* documents the case of a European agency creating and operating an AI-created influencer who makes approximately $11,000 per month.[M] Other consumer industries, such as mature entertainment–seen as a bellwether of technology adoption and proliferation– are adopting AIGC to develop original content to monetize for-fee services.[M] Driven by profit, AIGC technologies will almost certainly increase in sophistication and undetectable characteristics.

Advanced AIGC innovation will almost certainly continue to proliferate to ordinary users. Instructional videos are abundant online for creating hyper-realistic personas, demonstrating AIGC's production potential and diffuse nature.[M] Basic level users, within minutes, can create videos of themselves speaking in other languages and produce very realistic podcasts that develop a perception of authority (See Figure 2).[H] As more AIGC is available online and can be created quickly, AIGC detection will almost certainly become more critical to businesses and national security.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were reliable and tended to corroborate one another. NeptuneAI and Perplexity were used, and ideas from the results were utilized in further research. Perplexity suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information and advancements in AI technology and algorithms.

Author: Mr. Tom M. Jackson

By 2027 Adversaries Using Deepfakes Likely to Target SCADA Systems

Executive Summary

It is likely (55-80%) that by 2027 U.S. adversaries will use Artificial Intelligence Generated Content (AIGC), specifically deepfake videos, to attack U.S. civil infrastructure targeting critical nodes operating on Supervisory Control and Data Acquisition (SCADA) systems. SCADA systems are the backbone for industrial control systems, some include water treatment, power generation, transportation command and control and oil and gas. Due to advances in AIGC that can cause SCADA software disruption, override or even shutoff, it is probable (55-80%) adversaries will target these systems by 2027. Despite the U.S. Department of Defense recognizing the importance of cyber defenses, it is unlikely (20-45%) that defenses will protect SCADA systems.

Discussion

Supervisory Control and Data Acquisition (SCADA) systems are the "brains" of an industrial control system and have transitioned from standalone entities to now being connected to the Internet for more efficient communication of data.[H] See figure 1 for an example of a SCADA schematic.[H] The transition to a highly interconnected network has made SCADA more vulnerable to various cyber-attacks, given that security approaches provided by IT-based systems are not efficient enough to deflect the risks and threats.[H]

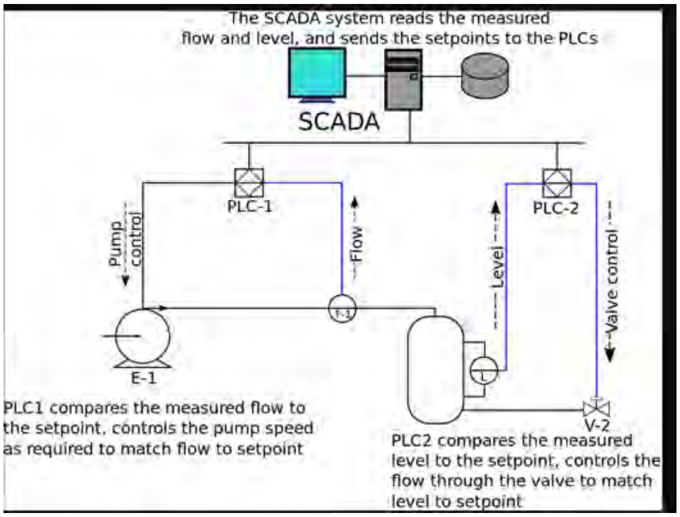

The SCADA system reads the measured flow and level, and sends the setpoints to the PLCs
SCADA
PLC-1
PLC-2
Pump control
Flow
Level
Valve control
F-1
L
E-1
V-2
PLC1 compares the measured flow to the setpoint, controls the pump speed as required to match flow to setpoint
PLC2 compares the measured level to the setpoint, controls the flow through the valve to match level to setpoint

Figure 1: Example of a SCADA schematic, referenced in the paper "Cybersecurity and the U.S. Energy Grid," by Nikhil Partasarathy, submitted as coursework (Stanford University, Fall, 2016)

Artificial Intelligence Generated Content (AIGC) includes a wide array of automated technology using algorithms that are 'trained' without human cognition.[H] One such area within AIGC that adversaries could leverage to their benefit is the efficiency of codewriting, and use of false data to target existing civil infrastructure nodes.[H]

Using AIGC advances, an adversary would write false code, having already bypassed any security systems and obfuscating attribution, cause a valve shutoff or a system override.[M] This situation would likely cause mass domestic panic for the U.S. government and require subsequent federal emergency response actions. Deepfake videos could have embedded malicious code in them for SCADA system users to download and initiate the attack. This exact scenario occurred in 20012 against the Saudi Aramco, the largest energy company in the world, when an image of a burning American flag had been downloaded on a computer and initiated a malware attack.[M] Targeted attacks against critical civil infrastructure would be an ideal means to disrupt an adversary in gray zone activities[i] and even continuing throughout full scale conflict, as seen in the existing Russia-Ukraine war.[H]

Despite the U.S. government's calls for improved security in the cyber domain, with its recent publication of the US National Cyber Strategy, as well as passing numerous bills of legislation focused on enhancing cybersecurity[M]; SCADA systems are still unprotected. The computer networks cannot be completely pulled off the networks to complete security upgrades and must rely on patch upgrades, which often can have only incremental, disparate, and limited protection measures. Moreover, due to its cost effective nature and ability to be operated remotely, SCADA systems will continue to be the preferred vendor of choice for managing industrial bases, remaining highly vulnerable.[M]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information, such as the SCADA systems being replaced or specific tradecraft being identified that enables better protection for SCADA systems.

Author: LTC Katherine M. Ogletree

Exponential Growth in AI Enabled Cyber Phishing Highly Likely to increase by 2028

Executive Summary

It is highly likely (88-92%) that the integration of Artificial Intelligence Generated Content into Cyber Phishing will drive exponential growth in cybercriminal activity, resulting in more than $90 billion in loss by 2028. Implementing international norms that inhibit the use of malicious AI will probably be the most effective means of countering AI-enabled cyber phishing.

Discussion

The Federal Bureau of Investigation lists the top five most prevalent categories of cybercrime activity as Tech Support, Extortion, Non-Payment / Non-Delivery, Personal Data Breach, and Phishing. The growth in the Phishing category alone exceeds the growth in all other categories combined. Augmenting current Phishing techniques with AIGC is highly likely to generate exponential growth in what is already a significant problem.[H] Potential AIGC-assisted attacks will almost certainly (95-99%) include Deepfakes to counter facial recognition software.[M] Criminal organizations will also us AI Powered Password Cracking to compromise business and government email accounts.[H] Another probable malicious use of AICG is poisoning legitimate AI data pools to make subtle changes to the parameters that inform AI's [M] Regardless of approach the goal will be to gain access to networks and will very likely include some form of advance phishing enabled by AIGC.

An emerging area of concern and area of probable growth is in Social Engineering in the cyber domain.[H] AI's are highly effective at spotting patterns in behavior and understanding how to convince people that videos, phone calls, or emails are legitimate by analyzing patterns and mimicking idiosyncratic tells at machine speed.[H] As a result, once Artificial Intelligence Generated Content (AIGC) is incorporated into the currently highly successful Phishing, if unchecked, will likely result in exponential growth in the field.

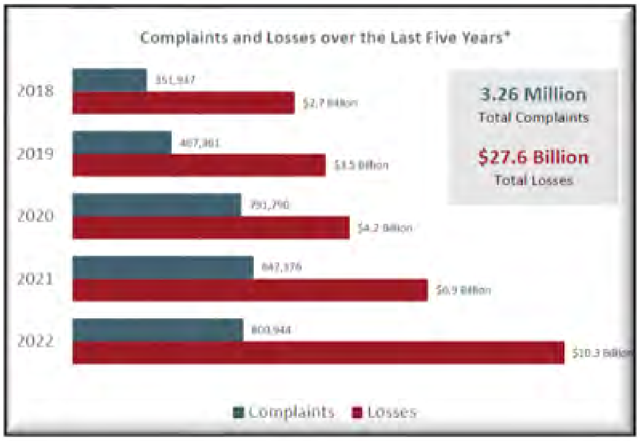


Figure 1. Financial losses due to cybercrime. Go to https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

Global criminal organizations are successfully conducting cyber attacks for profit at an increasing rate (see figure 1).[H] Domestic U.S. cybercrime in 2022 constituted $10.3 billion in losses, conversely the number of interdictions is decreasing. [H] Not only is the frequency of cybercrime increasing, but the monetary loss is increasing at a similar rate, averaging an annual growth of financial loss at 72% over the last five years. [H] At the current growth rates losses will exceed ~$92 billion by 2028 at the current growth rate The net result is that cybercrime is becoming more effective over time.

While the volume and draw of cybercrime increases, so does the sophistication with the application of AI and is likely to accelerate the pace of cybercrime.[H]. With the financial motivation to do so already evident, incorporating AIGC into Phishing attacks will likely increase the speed and volume of attacks exponentially. Incorporating AIGC into other forms of cybercrime is also likely (55-80%) to result in increased speed, frequency and sophistication of attacks.

A key factor underscoring this forecast is the domestic and international efforts to regulate AI use, which are unlikely (55-80%) to take effect in the next 4-5 years.[M] Given that the vast majority of cybercriminals reside outside the borders of the state they are active in to avoid arrest makes international cooperation a critical nexus between law and policy. To address the issue domestically, the Biden Administration has already put Executive Order 14111 into effect and will follow with legislation that governs the malicious use of AI. Additionally, the European Union and China are following suit with similar measures to safeguard against the nefarious use of AI.[H, M] The practical implementation of these measures will not take effect for the next few years but has a moderate probability (55-80%) of influencing the use of AI in support of cybercrime. [M]

Analytic Confidence

The analytic confidence for this estimate is *high*. Sources were highly reliable, tended to corroborate one another, and were generally authoritative and well-informed. There was adequate time, but the analyst worked alone and did not use a structured method. The likelihood of realizing this forecast depends on a few key variables, specifically the practical implementation of AI controls that result in effective international norms that enable the attribution and prosecution of cybercriminals.

Author: COL Robert M. Richardson

Uneven Implementation Will Likely Hinder Artificial Intelligence Generated Content Mitigation by 2033

Executive Summary

Public and private sector efforts to mitigate artificial intelligence generated content (AIGC) disinformation will likely (55-80%) have uneven effects on human perceptions of truth and misinformation by 2033. Government regulations will very likely (80-95%) temper the risk of AIGC but will not completely mitigate AIGC effects. Private sector technical means of AI detection and labeling will very likely alleviate AIGC misinformation but will have uneven distribution across media platforms. Media literacy and education efforts will almost certainly (95-99%) mitigate the effects of AIGC on the psycho-cognitive domain,[3] but uneducated groups will remain susceptible to AIGC's effects. Despite education, human perceptions of truth will almost certainly enhance, not mitigate, the effects of AIGC, creating vulnerabilities for exploitation by a foreign intelligence service. Though humans have been exposed to AIGC since 2019, there is a roughly even chance (45-55%) that AIGC exposure will provide an inoculation effect on the psycho-cognitive domain by 2033.

Discussion

Regulations and policies, such as restricting the creation of deepfakes, will likely mitigate AIGC's disinformation risk by 2033. The effects will be unevenly applied to the psycho-cognitive domain without an integrating strategy to overlap with technical and educational measures (See Figure 1).[H, H, M] The United States Intelligence Community's (IC) Human Intelligence (HUMINT) training and reporting framework is an example of applied measures that would almost certainly mitigate AIGC effects.[H] The United States'

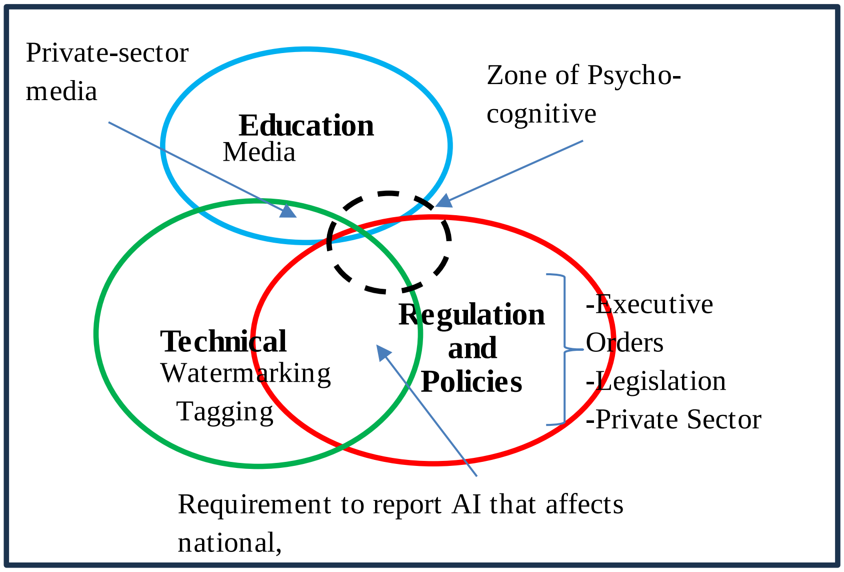


Figure 2: Venn Diagram connoting the current intersection of regulation, technical, and education measures necessary to mitigate effects on the psycho-cognitive domain. Dotted circle indicates the where the measures would need to overlap to have the best effect on the psycho-cognitive domain.

[3] On 8 November 2024, during the pre-meeting for the Terms of Reference, LTG Laura Potter defined effects on the psycho-cognitive domain as those things that would influence the will of a service member not to do a needed action or the will of the American people not to support the war fight.

open source intelligence (OSINT) collection apparatus, which has a regulatory framework as disparate as the current AI regulatory environment,[H] will almost certainly be susceptible to AGIC effects because of a lack of unified policy.[M] As an example of regulatory and technical overlap, Executive Order 14110, released in October 2023, addresses the need for the "safe and secure" use of AI and requires companies with AI systems that pose national security, economic, or health risks to notify the federal government.[H] The Department of Commerce, through its *National Institute of Science and Technology*, is implementing guidelines for accountability and AI risk.[H, M] In March 2024, the United States and the United Kingdom signed a bilateral agreement on AI testing and cooperation is a policy measure.[H] In March 2024, Tennessee passed a law to protect artists from AI reproductions of their voices and content, but it is uncertain if other states will adopt similar laws.[H] Indicative of the market approach, Tennessee's law is specifically targeted to protect its music industry, not to protect the psycho-cognitive domain of its citizens. [H] Together, these policies, and regulations address components of AIGC but do not confront the totality of AIGC's effects, diluting the efficacy of these efforts.

Private industry technical measures will very likely provide consumers with enough information to assess if the content they are viewing is AIGC, but not in a unified or integrated manner that would mitigate all of the effects of AIGC. By 2028, humans and AI detection models will almost certainly be unable to detect hyper-realistic AIGC, fueling an arms race of creation and detection. [M, M, M, H] Statistics firm *Precedent Research* projects that by 2032, the AI market will increase to approximately $2.5 trillion (See Figure 2).[H] The increased AI market share will very likely drive the demand for content certification as companies hasten to guarantee their products are genuine and can be trusted.[M] Recognizing that users on its platforms are susceptible to AIGC, and to create momentum for industry standards, social media company *Meta* recently announced it is labeling AIGC on its sites.[H] Companies, such as Intel, are partnering with organizations like The *Coalition for Content Provenance and Authenticity* (CCPA) to help

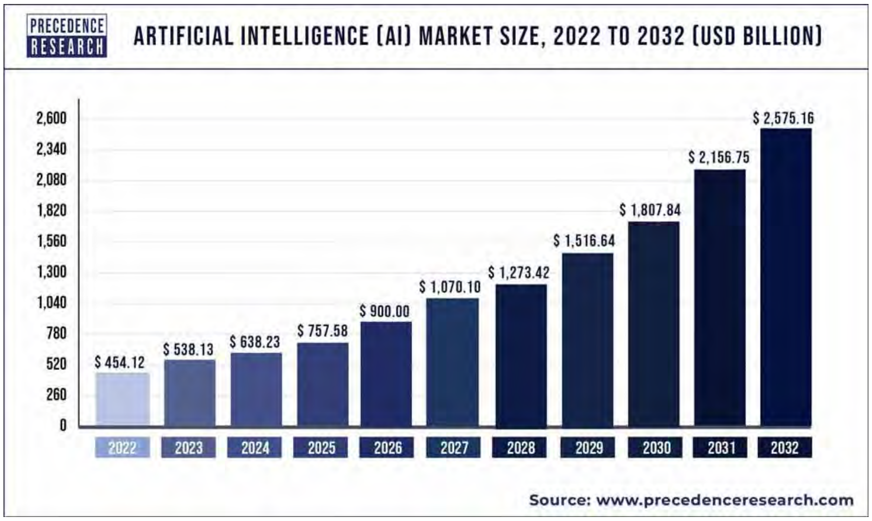
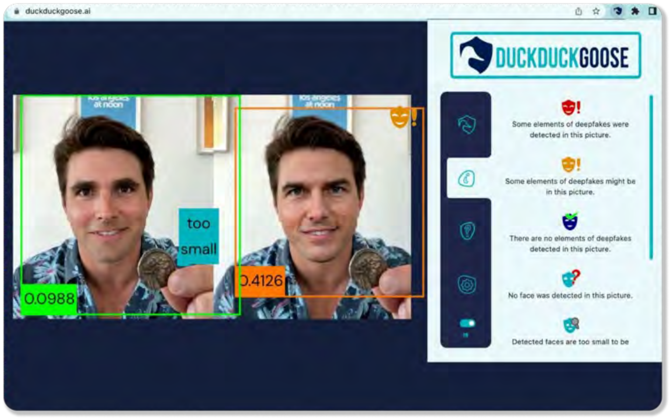


Figure 3: AI content marketing is projected to increase online by 25 percent. Click on the graph or go to https://www.precedenceresearch.com/artificial-intelligence-market Source: Precedent Research.



Some elements of deepfakes were detected in this picture.

Some elements of deepfakes might be in this picture.

There are no elements of deepfakes detected in this picture.

No face was detected in this picture.

Detected faces are too small to be

Figure 4: AIGC browser extension chart. Click on the chart or go to https://www.duckduckgoose.ai/deepfakeproof. Source: DuckDuckGoose

establish industry norms and frameworks for AIGC.[H] Other private industry approaches include AIGC detection companies like *OriginalityAI* and *Undetectable*, which are profit-driven AI detection services.[M, M] Due to the inundation of AIGC online, by 2028, search engines, such as *Google*, *Bing*, *Firefox*, will very likely provide AIGC monitoring, like browser extension *DuckDuckGoose*'*s* real-time AIGC monitoring[H] (See Figure 3) to alert users to likely AI content. Another method is to "cryptographically sign" AIGC, like AI-detection company *TruePic*.[M, H] These measures are essential to mitigate the effects of AIGC[M] but when applied in a decentralized manner, and without the force of legislation, it decreases their overall effectiveness.

Educational programs integrated with technical and regulatory measures will almost certainly mitigate the impacts of AIGC on the psycho-cognitive domain by 2033. Stanford University researcher Dr. Herb Lin, former Under Secretary of Defense Michèle Flournoy, and Professor of Psychology at Cambridge Sander van der Linden all note that education and AI familiarization are critical to preventing the effects of AIGC on the psycho-cognitive domain.[H, H] The *National Artificial Intelligence Advisory Committee*, in their 2023 *Recommendations: Enhancing AI Literacy for the United States of America* determined that education and AI literacy were necessary at a national level to confront AIGC misinformation.[H] Countries such as Finland have adopted whole-of-government approaches to mitigating the effects of AIGC through media literacy programs, resulting in greater resilience to disinformation.[M] Without a unified approach, the uneven application of media literacy in the United States will very likely create populations with differing levels of resilience.[H]

Regardless of educational measures, AIGC will almost certainly continue to challenge perceptions of truth through 2033.[H, M] This subjectivity, combined with AIGC, creates an environment wherein consumers will likely trust content if it aligns with their version of the truth, whether or not they know it is AIGC.[M] A foreign intelligence service will almost certainly exploit this vulnerability through the use of adversarial AI (AAI) chatbots to target military populations by 2033.[4], [H] Adopting processes, such as education and a "zero-trust" mindset [5] will very likely mitigate the effects on the target groups.

There is a roughly even chance (45-55%) that the proliferation of AIGC will reduce AIGC's disinformation effect on the psycho-cognitive domain by 2033. University of Notre Dame computer scientist Walter Scheirer found that because deepfakes originated in 2019, information consumers have been used to viewing and analyzing AIGC.[M] Researchers in the United States IC also assessed that prolonged exposure to misinformation has an aspect of diminishing returns.[H] This would indicate that the proliferation of AIGC misinformation, even if undetectable by humans, would likely have a diminishing impact by 2033. Conversely, researchers studying prolonged exposure to deepfakes found that deepfake exposure did not create a greater awareness for AIGC and had the opposite effect of increasing vulnerability.[H] The study also found that K-12 and college-age students were more capable of detecting AIGC than adult educators (See Figure 4), indicating that traditional considerations, such as education level, to lower vulnerability are unlikely (20-45%) to mitigate AIGC's effects.

Analytic Confidence

The analytic confidence for this estimate is *High*. Sources were reliable and tended to corroborate one another. *PerplexityAI* suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Given the lengthy time frame of the estimate, this report is sensitive to changes. Future quantitative studies will likely provide new information about the susceptibility of specific populations to AIGC. The information in this report is derived from open source, and sensitive reporting will likely contain more

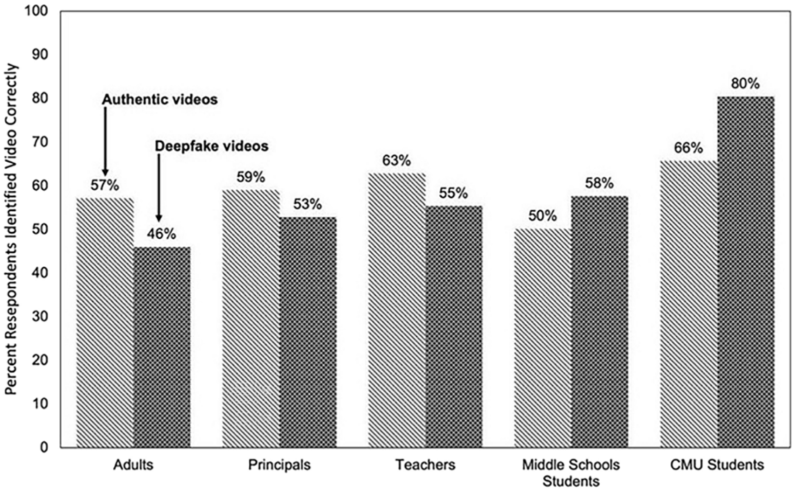


Figure 4: Survey chart of adults' and students' ability to recognize deepfakes. Groh et al. at the M.I.T. Media Lab found that humans were 65 percent better at detecting AIGC than AI models, which matches this study's results for Carnegie Mellon University students.[M] Click on the photo or go to https://www.nature.com/articles/s41598-023-39944-3#citeas. Source: Scientific Reports.

[4] Dr. Edmon Begoli, Artificial Intelligence Effects on the Psycho-Cognitive Domain, Teams Interview, February 29, 2024.

[5] Alper Kerman et al., "Implementing a Zero Trust Architecture," March 17, 2020, https://csrc.nist.gov/pubs/pd/2020/03/17/implementing-a-zero-trust-architecture/ipd. *National Institutes of Science and Technology:* Kerman et al. defined a "zero-trust" approach as that which "…removes the assumption of trust from users and network.

information on the propensity for foreign intelligence services to use LLMs and machine learning.

Author: Tom M. Jackson

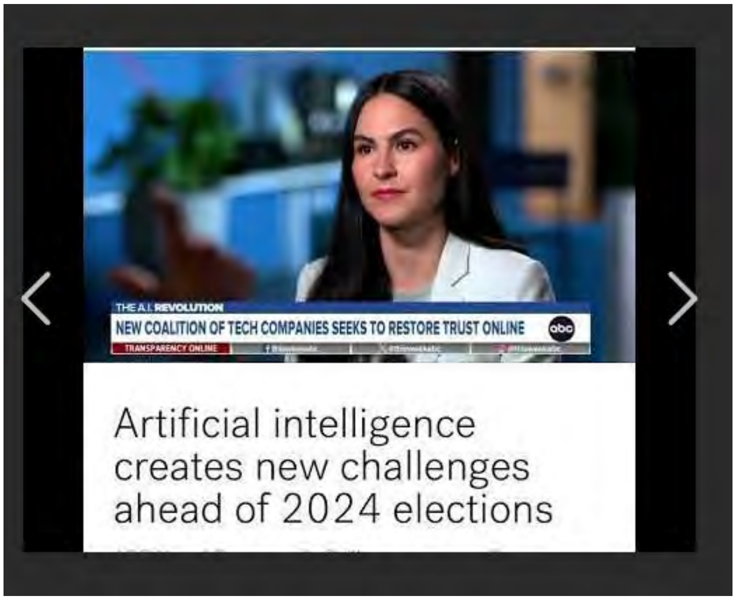
Most Countries Will Have Incorporated "Digital" Psychological Operations in How They Fight by 2029

Executive Summary

Rapid advances in deepfakes and Artificial Intelligence Generated Content (AIGC) and the adversaries' need to dominate the information domain during conflict make it almost certain (95-99%) that most countries will incorporate "digital" psychological operations in how they fight by 2029. Existing successes in "digital" psychological operations are due to adversaries having ease of access to a populace that is more technologically reliant than in years past. Despite AIGC vulnerabilities, such as hallucinations or inaccuracies, adversaries have ample opportunities within the next few years to refine their cyber skills. Events such as multiple global election cycles will allow trial and error, with adversaries learning how to mask their digital footprint better with every incident.

Discussion

Digital psychological operations, for the purposes of this paper, are activities which have the same end state as historically defined psychological operations but occur over technology or cyber platforms.[H] In recent years psychological operations have been synonymous with military information support operations (MISO), using technology to influence perceptions, shape behaviors, and guide decision-making processes among target audiences.[M]


THE A.I. REVOLUTION
NEW COALITION OF TECH COMPANIES SEEKS TO RESTORE TRUST ONLINE
abc
TRANSPARENCY ONLINE

Artificial intelligence creates new challenges ahead of 2024 elections

Figure 1: AIGC creating new challenging ahead of the US elections. Click on picture or go to: (25) Artificial intelligence creates new challenges ahead of 2024 elections ABC News' "This Week." - YouTube

Over ten years ago, US Central Command was engaged in a difficult counter terrorism fight.[M] Operation Earnest Voice used technology as the delivery method to reach a population, using fake online identities to influence internet conversations and spread strategic communication.[H] Since then, globalization has brought more connectedness, and citizen journalism[6] has become commonplace.H This is especially true with generations who grew up in the rise of the information age (25-34 year olds), even more so for those who grew up in the world of the social, participatory web, using such platforms as TikTok or Instgram (18-24 year olds).M The rise of AIGC has directly impacted digital psychological operations. Countries who may have been just beginning to use technology years ago can now do more operations, do them faster, and produce more catastrophic effects, with even less attribution.

In 2024, the possibility for AIGC to influence and spread mis and disinformation is likely the highest it has been since its inception due to the high number of global elections.M Figure 1 details how AIGC might be used as a digital psychological operations tool, creating new challenges ahead of the 2024 US elections.M

Voting countries will have digital footprints, which can exploited or affected, allowing for nearly an unlimited amount of practice to hone their cyber skills.M How U.S. adversaries fight in 2029 will have integrated digital psychological operations, with incorporations from lessons learned after many usages. Just like with any new fielding of adversarial equipment, the U.S. will need to answer priority intelligence questions, such as, what is the effect of the new technology and can it be countered? It is likely that by 2029, any digital influence will be easily shaped due to cognitive bias and the misinformation effect.H The misinformation effect refers to a type of memory impairment caused by introduction of (intentional) misleading information. M

Despite AIGC's vulnerabilities and weaknessM, such as hallucinations or inaccuracies, countries can still use even the weakest or slowest tools to successful employ digital psychological operations. In fact, that is precisely the point of psychological operations, to cause uncertainty, fear, doubt, and exploit these vulnerabilities so the limitations of AIGC in fact work toward the advantage of the psychological operations endstate.M

Russian involvement in the US 2016 election was successfully at creating this anxiety and doubt.H Moreover, in the run up to the Taiwan election, China has been accused of using AIGC to target the Taiwan People's Party founder and presidential nominee.M While China's desired opponent did not win the election, the amount of influence operations that was present may have ultimately pushed smaller regional countries closer to China, having abdicated from historical allegiance to Taiwan. M

[6] Citizen journalism defined by Britannica (online) is journalism that is conducted by people who are not professional journalists but who disseminate information using Web sites, blogs, and social media. The term and the practice crystalized in South Korea, where the online entrepreneur Oh Yeon-ho declared in 2000 that "every citizen is a reporter." (https://www.britannica.com/topic/citizen-journalism, accessed February 1, 2024)

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information, such as AIGC detection software becoming commonplace or global regulatory guidance on cyber warfare taking effect.

Author: LTC Katherine M. Ogletree

China Very Likely to Augment Cyber-Attacks With Artificial Intelligence-Generated Content by 2029

Executive Summary

China is very likely (82-85%) to augment national cyber operations with AIGC to increase sophistication and volume of attacks by 2029. China is also already involved in establishing international norms and controls, which will be the primary delimiting factor. Regardless, China's malicious use of AIGC will likely (60-65%) outpace protections for the near future.

Discussion

China already has a robust and capable national cyber capability launching at least 13 known attributable attacks targeting sister nations in the first qarter of 2024.[H] Nine of the recorded attacks were without data exploitation, while four of them ended up with data being misused. Three of the cyberattacks initiated by China in the measured year, contained theft of data. [H] (see Figure 1) Further, China has already begun to target states with phishing attacks designed to gain access to networks, steal data, or manipulate perceptions.[M] China's focus on phishing attacks indicates that the priority is gaining access to networks and data for future exploitation.

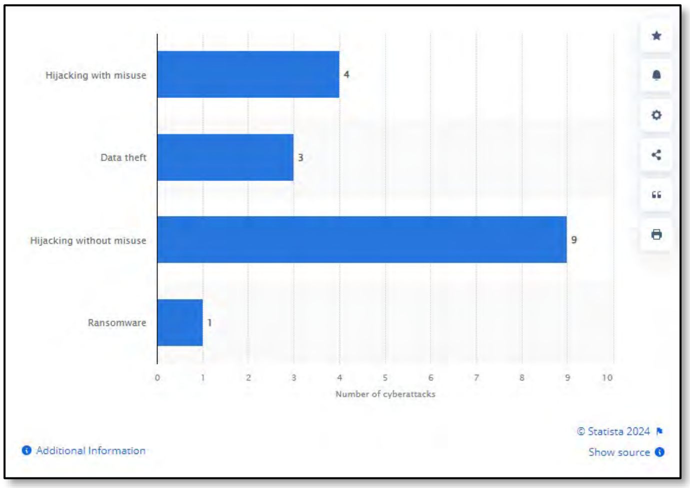


Figure 1 Main Types of Cyber Incidents with a Political Dimension Launched by China in 2023. Click on picture or go to: https://www.statista.com/statistics/1428527/china-launched-political-cyberattacks-types/#:~:text=In%202023%2C%20cyber%20threat%20actors,year%2C%20contained%20theft%20of%20data.

To date, China's cyber campaign is limited by the number of trained cyber personnel and access to networks. The introduction of AI will very likely (80-95%) allow Chinese cyber operators to shift to guiding AI tools increasing the number and scale of attacks. [M] Combine an experienced cyber team with the speed and power of AI and it becomes apparent that China will be capable of launching sophisticated cyber attacks on nations at an alarming rate within the next three to five years.

The only factor preventing China's use of AI Generated Content (AIGC) at the state level are the tale-tell continuity, distortion, or background flaws of current AIGC tools. [M] As AI tools

improve this limitation will fade. China is already making strides to incorporate AIGC into its ongoing cyber campaign. In 2023, the National Security Agency warned that Beijing started using AI to disseminate propaganda via a fake news channel.[M]
By augmenting state level phishing attacks with AIGC China will likely increase the volume and sophistication of attacks from a few a year to potentially, hundreds.

China is also learning from cyber criminals and has begun to augment its cyber capability with Artificial Intelligence (AI) increasing volume and sophistication of attacks.[H] China's cyber criminals are continuously developing new techniques to include incorporating AIGC to increase the sophistication of phishing attacks.[M] In 2021, Chinese authorities found that criminals had created AI-generated facial videos from scraped internet photos to illegally sign up for online payment accounts.[H]

There are two key limiting factors that will likely affect this forecast. The first is artificial. Most states, to include China are actively pursuing AI regulation. China's stated goal is to ensure "healthy development and standardized application" of AI services that can create text, videos, voice, and images. [M] China's seeks to shape international norms for its own benefit. However, if enough regulatory governing bodies agree on norms and guidelines, then the result would be a naturally self-regulating environment where participating parties regulate themselves.

The second factor impacting this forecast is technical. Using AI to detect AIGC is emerging but still nascent. It will likely be 3-4 year before an AI will reliably identify AIGC.[H] This means that the reliability of detection tools is already falling behind AIGC capability and is unlikely to reliably keep pace.[M] The use AI to detect AIGC is a growth area, but at present, the means to identify AIGC resides primarily in the human domain. As a result, China's use of AIGC for malicious means will almost certainly remain ahead of competitor states cyber protection.

Analytic Confidence

The analytic confidence for this estimate is *high* given that technical capabilities to conduct sophisticated cyber phishing attacks and AIGC capabilities already exist and only need to be combined. Sources were generally reliable and tended to corroborate one another. There was adequate time, but the analyst worked alone and did not use a structured method. This report is sensitive to change due to the pace at which international regulation is implemented.

Author: COL Robert M. Richardson

Artificial Intelligence Models Will Likely be Unable to Distinguish Artificial Intelligence Generated Content by 2033

Executive Summary

Technical advancements in Artificial Intelligence Generated Content (AIGC) make it likely (55-80%) that AI detection models, such as GPTZero and OriginalityAI, cannot detect all AIGC after 2028.[H] Despite efforts to develop AI detection mechanisms, AI detection will almost certainly (95-99%) not keep pace with AIGC advancement and creation.

Discussion

The evolution of AI algorithms and the creation of distinctly new AI algorithms will almost certainly outpace AI detection models, such as Intel's "Fake Catcher"[H] and for-profit companies like OriginalityAI.[M] Of the two commonly used AIGC algorithms, generative adversarial networks (GAN) and diffusion-based models, diffusion-based models are more capable of producing higher quality AIGC that would be undetectable by AI.[H] Diffusion differs from GANs in that a GAN has an adversarial algorithm that competes to produce a better outcome.[M] A diffusion model learns to create the image from "noise," creating a more detailed product (See Figure 1),[H] and diffusion models will very likely (80-95%) create more realistic and natural content than a GAN by 2028.[H] Because AI detection models look for characteristics of AI generation, such as discrepancies and alignment, [M], [M] the level of hyper-detail of diffusion models will almost certainly be undetectable by AI models by 2033. Notably, a distinctly new AIGC 3D model, Neural Radiance Field (NeRF), is promising to impact AIGC exponentially with hyper-realistic three-dimensional modeling from two-dimensional photographs. In 2020, a NeRF-generated model took several days to produce AIGC. In 2022, researchers at NVIDIA created a model in less than a minute.[M] While NeRFs take a large amount of computing power, it is almost certain that by 2028, NeRFs will be the next generation of commonly available AI models like GANs and diffusion.

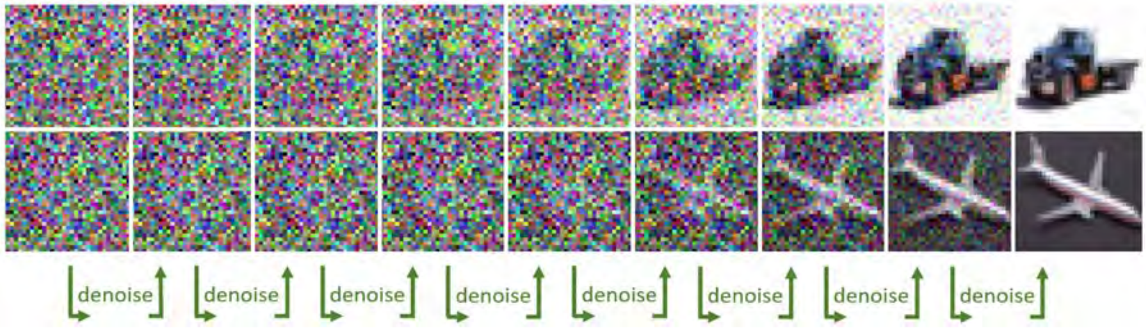


Figure 1: Diffusion AI produces an image out of "noise." Source: NVIDIA. Click to watch a video of a diffusion model tutorial, or go to https://www.youtube.com/watch?v=lyodbLwb2lY. Source: YouTube.

AI models alone will almost certainly not be sufficient to detect AIGC by 2033. Groh et al. at the Massachusetts Institute of Technology (M.I.T.) Media Lab established that humans were better able to detect AIGC than AI models due to the characteristics of how humans interact with images and videos, such as social cues. [H] However, AI detection models better identified grainy, dark, and odd-angle videos than human participants.[H] The inference from this research is that the development of detection models will very likely need to incorporate human-AI augmentation.

Detection models and algorithms responding to AI advancements will almost certainly follow, not lead, the creation of new AIGC methods. The monetization of AI encourages content creators to learn how companies like Google detect AIGC and use that information to produce higher-quality detection-proof content.[L] While detection algorithms are becoming equally sophisticated, they fuel a back-and-forth “arms race” of content creation and detection.[H, M] The decentralized and unregulated nature of AI encourages the advancement of AI that defeats detection as rapidly as the creation of detection models. In their off time, a Stanford University student produced a reliable AI app for generating college-grade essays, which was downloaded over 350,000 times in the first week. The app was specifically designed to bypass AI plagiarism detection.[M] Demonstrating the rapid evolution of AIGC technology, as recently as December 2023, researchers from Princeton and Carnegie Mellon created a new type of model, the structured state model, and assessed it would be five times faster than the transformer model, which is used for applications like ChatGPT, and has the potential to replace transformers and GANs.[M]

AIGC detection errors will likely enhance AIGC misinformation. The plagiarism detector GPTZero incorrectly identified legitimate text, such as the United States Constitution, as AI-generated, a form of AI-generated misinformation. Other detection models have similarly incorrectly attributed AIGC.[H] This will very likely sow mistrust of an AI model meant to provide detection assistance.[M] Former Undersecretary of Defense for Policy Michèle Flournoy understood this about AI when she stated in Foreign Affairs, “Without proper safeguards, AI models could cause all kinds of unintended harm. … policymakers will have to implement better approaches to accelerating adoption as well as ensuring safety.”[H]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were reliable and tended to corroborate one another. The analyst used PerplexityAI, and ideas from the results were utilized in further research. Perplexity suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to changes in private-sector policies and government regulations. Specifically, the technological acceleration of AI models and the rapid creation of distinctly new AI models add further sensitivity to changes.

Author: Mr. Tom M. Jackson

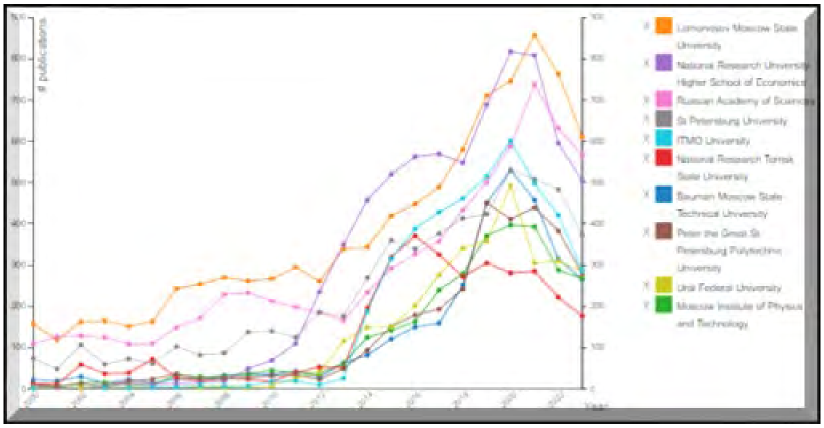
Russian Cyber Operations Highly Likely to Radically Evolve By 2030, Shifting from Brute Force Attacks to Cyber Espionage

Executive Summary

Significant national investment, integration of AI into cyber operations, and the advent of the Russian National Artificial Intelligence Center (RNAIC), and reaction to ongoing sanctions make it very likely (80-95%) that Russian AI-enabled cyber operations will radically evolve between 2028 and 2030 despite the war in Ukraine.

Discussion

Over the last five years the Russia's AI base of expertise has experienced considerable growth from 14 startups to 434 companies.[M] The academic research arena has experienced similar growth as indicated by a poll of academic publication on the topic with a drop off following the invasion of Ukraine.[M] (see Figure 1) These growth indicators do not include the governmental organizations that also have an AI presence, however, President Putin's policy of incorporating capability into all aspects of government is clear.[H] Further, with the advent of the RNAIC, Russia is likely to evolve its cyber capability from a brute force model to a significantly more refined model capable of cyber espionage by 2030.



Using publications as an estimative base, Russian AI research has grown by more than 1000% over the last 10 years. Click on the picture or go to: AI Strategies and Policies in Russian Federation - OECD.AI Source: Statista

While it is difficult to decern the real effect of Russian cyber, where Russian cyber does have an impact, it is generally limited or short-lived and is easily attributable.[H] Russia's unrefined cyber-attacks are as a result of poor organizational structure which undermines effectiveness. Further, the war with Ukraine drives Russia to contribute considerable cyber resources to cyber-attack and active defense.[M] However, with the eventual end of the war with Ukraine combined with a new organizational structure under the 2020 established RNAIC potentially addresses several key limiting factors.

Despite committing considerable cyber resources to Ukraine, Russia is still investing heavily across the AI industry. Russian investment in AI research has grown from approximately $16 billion to more than $115 billion annually.[M] Additionally, the RNAIC mission is to foster partnerships among the nation's leading state-run and private companies and universities.

This step marks Russia's first attempt to establish and organizational structure to synchronize and integrate AI into Russian state and private cyber operations.[H] Finally, Russian policy directs the incorporation of AI into every possible area of government, public, and private life.[M] While investment, technical, academic, and policy conditions may be set, Russia may still struggle to fully implement AI enabled cyber operations as effectively as the west.

The war in Ukraine is the primary limiting factor inhibiting Russian AI but is also serves as a testing ground for emerging techniques and tools including AI.[M] In 2023, approximately two-fifths of the approximately 4,000 Russian cyber operations were oriented against Ukraine, while over the same period, Russia was the target of one inbound cyber-attack for every six outbound cyber operations.[L] These numbers do not account for unreported state-level cyber operations so the actual number of cyber events are likely much higher.

In the two to three years following the close of the war in Ukraine, new organizational structure offered by the National AI Center combined with the emergence of new technologies will likely drive Russia to evolve its cyber capabilities in innovative ways. Russian backed cyber actors have already been detected adding large-language models, like OpenAI's ChatGPT, to their toolkit, often in the preliminary stages of their hacking operations.[M] The National AI Center will almost certainly enable Russia to augment its cyber capability with emerging tools and technologies across state, private and academic arenas as such Moscow has set conditions for an evolution in AI enabled cyber operations shift from brute force attacks to subtle and refined cyber espionage operations.[M]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. There was adequate time, but the analyst worked alone and did not use a structured method. However, this report is sensitive to changing conditions outside of the technical nature of the report, specifically political conditions between Russia and Ukraine.

Author: COL Robert M. Richardson

Social Media Deepfakes Unlikely to Change Adult Opinions Through 2033

Executive Summary

Adults in the United States are unlikely (20-45%) to have their minds changed from seeing deepfakes on social media sites through 2033. Theories explaining the belief of truth have evolved over time and truth has become less fact based and more feeling based. Studies show that people develop their version of truth from their personal history and develop cognitive biases which are difficult to change due to the way the brain operates. While adults are increasing the amount of news they receive from social media, they remain less malleable to have it affect their pre-conceived foundation of truth. Despite the youngest adults being the most trusting of news received on social media, it remains unlikely they will overcome their bias and change their mind due to deepfakes.

Discussion

There are five major theories concerning how human beings determine truth and they have evolved over time. The *Correspondence Theory* correlates truth to facts.[H] The *Semantic Theory* expands on the Correspondence Theory and uses a metamathematical formula to ensure the right facts are used to find truth.[M] The *Deflationary Theory* breaks away from mathematical models and sees truth as a literary convenience.[H] The *Coherence Theory* indicates people find something true when it does not conflict with what they believe to be true.[H] Finally, the *Pragmatic Theory* observes that people find something truthful if it is useful, and untruthful if they do not find it useful.[H] These theories demonstrate that truth shifted from a fact based idea of truth to truth becoming a method to further one's personal and conditioned reasoning.

Since theories of truth evolved to allow truth to be subjective and not entirely fact based, changing someone's mind is difficult with fact-based discussion. A Harvard Business Review study of business leaders revealed that sound arguments and good presentation are required to dissuade someone to achieve cognitive conversion.[H] The same study showed that no amount of logical or emotional argument was successful at changing a deep-set opinion that was based on upbringing, cognitive bias, or personal history.[H] Further studies show that

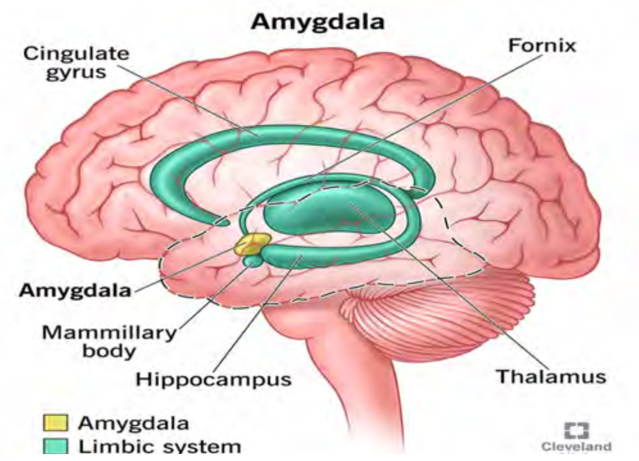


Figure 1 The amygdala is an almond shaped part of the human brain which detects danger. https://my.clevelandclinic.org/health/body/24894-amygdala

people do not accept new or uncomfortable information if it challenges their world view.[M] The part of the human brain which is triggered by conflicting information is the amygdala (See Figure 1).[M] The amygdala interprets changes in information as a threat and releases hormones to help the human body prepare for fear, fight, or flight.[M] The human brain is wired to resist new information which challenges cognitive biases. A Pew Research Center poll showed that half of adults in the United States get news from social media at least sometimes (See Figure 2), with Facebook being the most visited site.[M] Social media is largely responsible for the spread of deepfakes,[M] therefore half of adults in the United States are at risk of seeing deepfakes as news. The real risk is not someone believing the deepfake as real news; the true risk is not believing something is real due to being conditioned to disbelieve everything.[M]

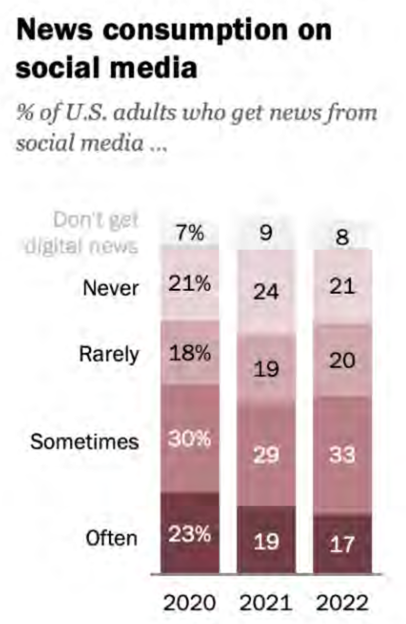


Figure 2 Pew Research results concerning news received from social media. *https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/*

The youngest adults, ages 18-29, are the leading news consumers on Instagram and TikTok.[M] Furthermore, this same age group is the most believing of news received on social media than any other age bracket of adults in the United States.[M] This group reports to have some, or a lot, of trust in what they see on social media.[M] This places the youngest group of adults at the most risk of being affected by deepfakes. They have had the least amount of time and experience to fully form their biases and are spending the least amount of time consuming non-social media news sources to corroborate or challenge information.

Analytic Confidence

The analytic confidence for this report is *moderate*. Sources were open-source documents which were generally reliable and academic sources were used where possible. Sources were corroborative and did not have major discrepancies in content. Adequate time was available to prepare this report, but the analyst worked alone without a structured method. Perplexity AI was used to help summarize theories. The lengthy timeframe of this estimate is sensitive to new information such as future studies showing the result of adults whom have been exposed to misinformation their entire life.

Author: LTC Charles Moss

By 2029 Countries Likely Noncommittal on AIGC Norms, with Training for U.S. Military and Countermeasures in Place Almost Certain

Executive Summary

By 2029, it is likely (55-80%) that most countries will not sign governance on Artificial Intelligence Generated Content (AIGC) behavior because it would restrict their own offensive capabilities in the information domain. Despite global rhetoric from leaders who have called for international norms or to increase the dialogue as a means to discuss AIGC usage protocols, it is not in their interest to sign regulatory and punitive governance. Understanding this, it is almost certain (95-99%) that the U.S. will appropriately take countermeasures against successful digital psychological operations, such as increasing training and cyber literacy.

Discussion

Even before the rise of AIGC, using digital psychological operations has been common practice as an irregular warfare tool.[H] Cyber warfare can be traced back as far as 2008 when Russia used it prior to its incursion into Georgia.[M] More recently, in January 2024, China used it to attempt to influence Taiwan's election.[M] While some tactics have changed throughout the years, the fact that cyber warfare has not been clearly defined and lacks clarity for its use on the spectrum of armed conflict is the interesting point.[M] remarkable. How is it that the international community has not set the distinguishable point.[H] This is likely (55-80%) due to the freedom and flexibility that remaining noncommittal on any international or regulatory order defining cyber warfare characteristics affords a country.[M] For instance, if a country uses cyber operations to attrite command and control of an adversary prior to armed conflict, but there is no international law governing these actions then there is potential for limited, if even any, consequences.[M]

Figure 5: General CQ Brown, the Chairman of the Joint Chiefs of Staff, discusses the risks of artificial intelligence in the military. Click on the picture or go to: https://www.bing.com/videos/riverview/relatedvideo?&q=the+use+of+artificial+intelligence+as+irregular+warfare&&mid=A429A270651D40777074A429A270651D40777074&&FORM=VRDGAR

Since the rise of AIGC, many countries have established and published frameworks calling for responsible behavior. However, these are usually voluntarily signed, singularly developed, and without stipulations of penalties.[M] Moreover, despite leaders stating they are ready to agree to and implement collective regulations,[M] they continue to use AIGC to conduct digital psychological operations to advance their own national interests.[M] If a country is not subject to an internationally signed agreement or bound by conventions or law then any advancements in that area can occur without global repercussions. Most recently this was demonstrated with Iran's nuclear weapons program and its withdraw from the Joint Comprehensive Plan of Action.[M] Specifically addressing AIGC and regulatory guidance, countries not bound to being a signatory of any document with global repercussions then have freedom to advance in AIGC much faster than those without regulations, and subsequently portends future use without repercussions. AIGC could truly maximize existing digital psychological operations during gray zone operations.[H] In figure 1, Chairman of the Joint Chiefs of Staff, General CQ Brown, details the risks of AI(GC), and how countries might use it for digital psychological operations.

Knowing the AIGC risks to national security, it is almost certain (95-99%) the U.S. military will take measures to protect and counter this threat. Efforts like using detection software, which can check the veracity of open source data, to include programs that identify videos and images as fake, should be expected as it can mitigate risk.[H] Additionally, professional military education (PME) and training to incorporate AIGC literary will be commonplace by 2029, specifically for intelligence analysts as they incorporate open source intelligence given their job requirements.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change. Actions that could affect this forecast include if agreed upon regulatory guidance is not punitive, or if the U.S. is in charge of the collective regulations.

Author: LTC Katherine M. Ogletree

Continued Investment By Industry in Identifying AI Generated Content is Very Likely to Continue Thru 2033

Executive Summary

Industry investment in detecting AI-generated content like deepfakes and bots is very likely (80-95%) to significantly increase through 2033. Advanced AI has made creating convincing manipulated content easier and more accessible, which will drive a proliferation of deepfakes and bots without mitigation efforts. However, if investment in promising detection technologies continues, it should help curb risks, although some fake content will inevitably slip through.

Discussion

Technological leaps in AI, such as generative adversarial networks (GANs)[H], have enabled new content manipulation techniques that require less expertise to deploy effectively. In 2019, the "DeepTomCruise" TikTok account demonstrated highly realistic Tom Cruise deepfake videos created by a single VFX artist, showcasing how advanced individual efforts have become.[M]

In response, a variety of detection technologies and approaches have emerged in recent years:

- Machine Learning Algorithms: Deep learning models like XceptionNet[M] have shown promising results in detecting deepfakes by analyzing visual inconsistencies. These algorithms can learn the patterns and features that distinguish between real and manipulated videos.[H]
- Forensic Analysis: Forensic techniques, such as analyzing inconsistencies in lighting, shadows, reflections, and facial movements, can be employed to detect alterations in videos. For example, the FaceForensics++ dataset is being used to train and benchmark deepfake detection models.[M] These methods often rely on detailed frame-by-frame analyses and comparisons with reference data to identify discrepancies.[H]

Social media companies come together to fight misinformation. https://www.ciobulletin.com/mobile/social-media-companies-fight-misinformation Source:

- Image and Video Authentication: Digital watermarking, digital signatures, and other authentication techniques can be applied to videos to ensure their integrity and prove their authenticity. These methods can help verify the source and detect any modifications to the video since its creation.[M]

- Collaboration with Social Media Platforms: Social media platforms are actively developing and implementing detection systems to identify and flag deepfake videos and disinformation campaigns. Major platforms like Facebook, Twitter, and YouTube are actively developing and implementing deepfake and bot detection systems.[M] They employ a combination of AI algorithms, user reports, and human moderation to detect and remove false content.[M]

- Data Verification and Fact-Checking: Fact-checking organizations and initiatives aim to verify the accuracy of information circulating online. Organizations like FactCheck.org are working to verify online information accuracy, with increased focus on AI-generated content.[H] They employ data verification techniques, source analysis, and expert judgment to identify and debunk disinformation.[H]

Several major social media platforms are investing in, and actively working on, detecting deepfake videos and disinformation campaigns. The following are some examples.

- Meta: Meta which includes Facebook and Instagram has been investing in research and development to detect and combat deepfakes and disinformation. They use a combination of AI algorithms, including machine learning and computer vision, to analyze content and identify potential manipulations. In the coming months, Facebook may explore labeling images that users post when it can detect indicators that are AI generated.[H] They also collaborate with third-party fact-checking organizations to verify the accuracy of information.[H]

- YouTube: YouTube, owned by Google, has been working to detect and remove deep fake videos and disinformation. They employ a combination of automated systems and human reviewers to identify and take action against misleading or harmful content. YouTube also relies on user reports to flag potentially problematic videos.[M]

- TikTok: TikTok implemented measures to combat deepfakes and disinformation on its platforms. They use AI-based algorithms to detect and remove manipulated content. TikTok also provides users with reporting options to flag false or misleading information.[M]

However, detection capabilities must constantly evolve as generation tools improve. More extensive datasets and computing power may allow subtle fake indicators to be found, but manipulators can leverage the same advances, fueling an AI "arms race." Generating fakes still faces fewer technical barriers than identifying them. Rigorous schemas for evaluating detection tools will be key for driving innovation.

While industry efforts should mitigate risks through 2033, the spread of some fake content is inevitable. Detection may lag around major events like elections when new techniques emerge rapidly. Longer-term, regulatory frameworks and digital literacy initiatives will also be critical. But profit motives and reputational concerns will continue incentivizing aggressive industry investment.

Analytical Confidence

Confidence level is *moderate* that industry will continue to invest in the detection of deepfake videos and disinformation over the next 5-10 years. This assessment is based on evidence regarding the direction of technology developments as well as unambiguous profit motives driving their investment despite ethical issues. Regulatory inability to constrain growth in the near term solidifies this outlook. However, some uncertainty exists around potential future actions by social media platforms if issues escalate further.

Author: CDR Robert V. Liberato

Human Intelligence Collection by the United States Unlikely Affected by Artificial Intelligence by 2033

Executive Summary

Human intelligence (HUMINT) collection within the United States Intelligence Community (IC) is unlikely (20-45%) to be affected by artificial intelligence generated content (AIGC) by 2033. The regulatory nature of HUMINT and the degree of human-in-the-loop within the collection cycle make it unlikely that AIGC will enter or remain in IC reporting through HUMINT sources. Though adversarial artificial intelligence (AAI) will likely (55-80%) be a challenge to HUMINT and HUMINT sources might provide AIGC, human collector training, and multiple humans in the collection cycle make it almost certain that AIGC would be detected.

Discussion

Regulations, policy frameworks, and established processes mitigate AIGC harm to HUMINT collection. Department of Defense Instructions (DoDI) require specific HUMINT collector training and certification, establishing professional standards for conduct.[H] HUMINT information then iterates through at least three more human touchpoints after collection: processing, analysis, and evaluation in the IC.[H] IC directives establish robust frameworks for assessing and processing HUMINT information.[H] Subject matter experts review collected information and turn it into intelligence reports.[H] Dissemination to the IC and evaluation acts as a "peer review" in that other IC analysts and consumers review the intelligence material and provide feedback.[H] In total, at least three human touchpoints review HUMINT-collected information in the intelligence cycle: the collector, the analyst, and various report evaluators. Even without additional AI familiarization training, it is likely (55-80%) that a human at one of these touch points would detect AIGC.

HUMINT is primarily a human-to-human interaction, and HUMINT collectors, with additional AI familiarization training, will very likely (80-95%) be able to attribute AIGC information, text, or images correctly. Dr. Herb Lin, an AI researcher at Stanford University, assessed that educating humans, not technology, would mitigate AIGC's harm to the IC. By qualitative testing, Groh et al. at the MIT Media Lab also determined that humans were marginally better at AI detection than selected AI models.[H] When augmented with AI detection models, the participant's detection rate increased by ten percent.[H] Former Under Secretary of Defense for Policy Michèle Flournoy considered education and exposure to AI critical for attribution and detection.[7] As a countermeasure to AIGC, the *World Economic Forum* recommends adopting a cybersecurity "zero trust" mindset.[H] This is based on a report from the *National Institutes of Science and Technology: Computer Security Resource Center* paper, which outlines that a "zero-trust" approach "…removes the assumption of trust from users and network."[H] An AI "mindfulness" program also increases the probability of AIGC detection using a psycho-cognitive approach that understands emotional triggers and takes a holistic view of the information.[H] An individual with this mindset acts as a "human firewall" for information, increases awareness of AIGC misinformation, and would lead to better detection.[H] HUMINT collectors and analysts trained in AIGC familiarization, mindfulness, and a "zero-trust" mindset would form a formidable mesh of detection mechanisms. Though a HUMINT source might unwittingly provide AIGC, such as misinformation collected from the Internet, a HUMINT collector trained in a "zero trust" mindset and AI mindfulness would very likely detect the AIGC.[M] If a HUMINT collector did not detect AIGC at the point of collection, the other human touch-points (processing, analysis, and IC review) would almost certainly (95-99%) identify AIGC.

Figure1: Microsoft digital briefing on threat actors using AI. Click on photo or go to https://s7d2.scene7.com/is/content/microsoftcorp/Cyber-Signals-Digital-Briefing-Issue_video_en-us. Source: Microsoft

Adversarial AI (AAI) will almost certainly challenge IC HUMINT reporting. A Department of Homeland Security report identified several AAI attacks potentially impacting HUMINT, such as generative AI, image morphing, and data poisoning[8].[H] As recently as February 2024, Microsoft researchers found that Iran, Russia, North Korea, and China were leveraging large language models for intelligence purposes, such as intelligence deception and intelligence-related content generation (See Figure 1 interviews of Microsoft security experts).[M] AAI-generated text that resembles sensitive adversarial information or AAI-generated images provided to HUMINT collectors through a double-agent[H] will likely pose a challenge to HUMINT collectors at the point of collection. Because the IC HUMINT enterprise has an established HUMINT collection and dissemination framework replete with human touchpoints, it is almost certain that AAI-provided information would be detected and

[8] Dimitri Kusnezov, Ph.D. et al., "Department of Homeland Security: Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats," Government (Washington D.C., June 2023), 13,20. Data poisoning is introducing data into an AI model to alter the model's outputs. Imagine morphing is using an AI model to combine two images into one composite image.

attributed to AAI. With additional AIGC familiarization training and a “zero-trust” mindset, it is unlikely that AIGC will enter IC reporting through HUMINT collection.

Analytic Confidence

The analytic confidence for this estimate is *high*. Sources were reliable and corroborated one another. Perplexity was used, and ideas from the results were utilized in further research. Perplexity suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information and how a foreign intelligence service would use AAIs for HUMINT operations. Some of the information on IC collection is classified, and assessments were made using publicly available information from the Director of National Intelligence.

Author: Mr. Tom M. Jackson

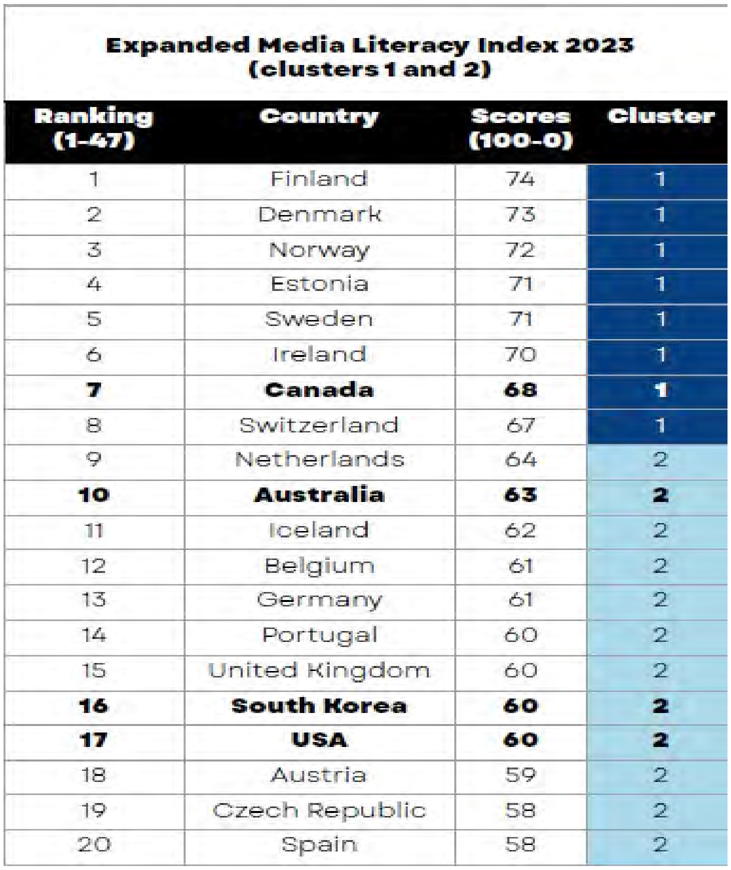
Likely the United States Government Will Prioritize Media Literacy Programs by 2033

Executive Summary

The rise of social media makes it easier than ever to access information, including information that is false or misleading. This is due to steady advancements in Artificial Intelligence Generated Content and the growing threat of disinformation. The need for further media literacy programs makes it likely (55-80%) that by 2030, the United States Government and Department of Education will develop a curriculum to target media literacy education to address the negative effects of disinformation and misinformation.

Discussion

The proliferation of disinformation threatens the very foundations of a healthy society by clouding truth and sowing confusion and discord among the population.[M] Disinformation has also been utilized to undermine elections, enable human rights abuses, justify conflicts, and challenge scientific consensus.[H] Education, especially at an early age, is an essential component to address disinformation and misinformation.[H] “Raising awareness among children and teenagers about misinformation is fundamental in preventing them from being tricked by falsehoods online,” said Aleksandra Wrona, marketing and education director of Pravda Association. “What we want to see in the future is a society resistant to misinformation”[H]



Expanded Media Literacy Index 2023 (clusters 1 and 2)

Ranking (1-47)	Country	Scores (100-0)	Cluster
1	Finland	74	1
2	Denmark	73	1
3	Norway	72	1
4	Estonia	71	1
5	Sweden	71	1
6	Ireland	70	1
7	**Canada**	**68**	**1**
8	Switzerland	67	1
9	Netherlands	64	2
10	**Australia**	**63**	**2**
11	Iceland	62	2
12	Belgium	61	2
13	Germany	61	2
14	Portugal	60	2
15	United Kingdom	60	2
16	**South Korea**	**60**	**2**
17	**USA**	**60**	**2**
18	Austria	59	2
19	Czech Republic	58	2
20	Spain	58	2

Figure 1 Expanded Media Literacy Index. Source: 2023 Open Society Institute Media Literacy Index https://osis.bg/wp-content/uploads/2023/06/MLI-report-in-English-22.06.pdf

A 2016 Stanford University study, showed that high school students had a hard time distinguishing between sponsored content and news articles or determining the potential bias of social media messages.[H] A 2018 MIT study on the spread of false information on X, at the time known as Twitter, found that fake news stories were 70 percent more likely to be retweeted than true news stories.[H]

The United States lags behind many other countries in areas that indicate effective media literacy education.[M] According to the latest report, Finland tops the media literacy index for the sixth time in a row, with the United States ranking 17 out of 47 countries (See Figure 1).[H] The index measures potential vulnerability to disinformation, with higher rankings and scores indicating better resilience of societies to the impact of disinformation and related phenomena.

Several factors suggest the United States government will likely prioritize media literacy education by 2030:

- The increasing sophistication and prevalence of disinformation campaigns threaten social stability and democratic processes. Left unchecked, these trends will worsen in the coming years.
- Successful models in countries like Finland demonstrate the efficacy of comprehensive media literacy programs in building societal resilience against disinformation. Adapting these approaches to the U.S. context is feasible with sufficient political will and resources.

- Educating the younger generation is a proactive, long-term solution compared to reactive policy measures. A media-literate populace can identify misinformation and approach online content with appropriate skepticism.[H]

- The costs of inaction, in terms of social discord and weakened democratic institutions, likely outweigh the financial and logistical challenges of implementing media literacy curricula.

While valid concerns exist around costs, federal overreach, and free speech implications, the urgent need to equip citizens with critical media skills will likely spur government action within the forecast timeframe. By 2030, without significant changes, civilians entering the military will almost certainly be susceptible to misinformation and require media literacy training. An informed populace is harder to exploit and more resilient to the negative effects of disinformation over time.

Analytical Confidence

Confidence level is *moderate* that United Staes Government will develop media literacy curriculums within the next 2-6 years. This assessment is based on clear evidence regarding the direction of media disinformation and misinformation developments and variations in the adaptability of international media literacy programs to the U.S. educational system, potential resistance from stakeholders, and the evolving nature of digital misinformation. Regulatory inability to constrain growth in the near term solidifies this outlook.

Author: CDR Robert Liberato

Social Media Industry Unlikely to Prevent Misinformation Through 2033

Executive Summary

Social media companies are unlikely (20-45%) to prevent the posting and spread of misinformation on their platforms through 2033. TikTok is designed to be addictive for young adults and is the easiest social media platform to use for manipulation. Facebook is attempting to ensure misinformation is removed but its process is slow. X has shown success in removing posts with misinformation All the referenced companies have policies concerning the removal of misinformation, but their policies are slow to react and the spread of misinformation has already occurred.

Discussion

TikTok's community guidelines allow the use of artificial intelligence generated content (AIGC) and other forms of digital technology.[H] These posts must be marked with a notification indicating it is synthetic or not real.[H] These guidelines specifically disallow deepfakes to be posted on private individuals, but deepfakes concerning public figures have more latitude.[H] TikTok's community guidelines do not give specific guidance on how deceptive or manipulative videos must be marked or how conspicuous the marking must be. Independent research shows that TikTok often lacks context and citations which people need to verify information.[M] This is of particular concern as younger generations have less awareness of fake news, when compared to older generations.[M] Generation Z is the most trusting of TikTok among adult users in the United States, at 65%.[M] A separate study showed further dangers with TikTok, finding it was the most addictive social media platform.[M] This study indicated TikTok's algorithm is intentionally designed to be addictive in an effort to drive users to specific videos.[M] Francesca Panetta, a multimedia artist and journalist, cited TikTok as a social media

Figure 1: Meta CEO, Mark Zuckerberg, testifies before congress concerning community guidelines. Video available at: https://www.youtube.com/watch?v=VDmeGQcpRLQ

platform which is designed to merge various media sources, making it useful for manipulation.[M] Panetta also questioned the platform's ability to perform fact checking.[M] The addictive nature of the platform, coupled with its corporate policies are unlikely to change due to the company's success in the United States. TikTok was estimated to have 192,000,000 North American users in 2023.[M]

Meta, the company which owns Facebook and Instagram, has similar rules regarding manipulated media, but are stricter. Meta has two criterion to determine if a video should be removed: a video is not detectable to be misinformation to the average person and the video was created by AIGC which appears to be real.[H] Further, Meta reserves the right to label content which, although does not directly violate community standards, if it is AIGC and has a high risk of deceiving the public on important matters.[H] Figure 1 shows Meta CEO testifying before congress concerning the platform's community guidelines. A report from George Washington University discovered that Facebook's policies are not effective at systemically removing misinformation.[H] This is due to the design of the platform, which is intended to build communities based on similar interests and ideologies.[H] A report from the United States Department of Homeland Security credited Facebook with being able to detect deepfakes but also found that the detection is slow.[H]

X, formerly known as Twitter, has a more direct policy concerning synthetic material. X's overview of its community guidelines concerning AIGC is "You may not share synthetic, manipulated, or out of context media that may deceive or confuse people and lead to harm ("misleading media")."[H] X also withholds its authority to label posts which contain misinformation.[H] The Cambridge University Press credited Twitter with actively taking steps to reduce misinformation on their platform.[M] The study revealed Twitter followed its policies of removing posts which were found to be potentially harmful if someone believed the post. [M] A separate study concluded that Twitter posts with misinformation spreads six times faster than truthful posts.[M] Misinformation is not the only issue that Twitter should be aware of. A 2021 study showed that X's user algorithm is similar to that of TikTok; it is intended to keep users within their echo chamber.[M] While X has shown success in removing posts which violate its user agreement, it is unlikely to create a mechanism which would perform fact checking prior to allowing a post to be seen.

Analytic Confidence

The analytic confidence for this report is *moderate*. Sources were open source documents where were found to be reliable and corroborated information. Corporate community guidance rules were sourced directly from the companies' websites. Academic sources were referenced where possible. Corporate policies are constantly changing and could have future impacts on this estimate. Adequate time was available to prepare this report, but the analyst worked alone without a structured method. The lengthy timeframe of this estimate is sensitive to new information such as updates to corporate policies.

Author: LTC Charles Moss

Artificial Intelligence Generated Synthetic Data Provides Novel Approach to Infiltrating Networks Between 2025 and 2030

Executive Summary

The use of AI generated synthetic data to train AIs is highly likely (80-95%) to pose an increased risk to network security due to poor data supply chain efficacy between 2025 and 2030 despite cyber protections.

Discussion

As AIs become increasingly complex and varied in use, the industry is moving to synthetic data to train AIs. Machine learning requires millions of samples to train a neural network.[L] In training an AI, it is often technically infeasible to create every possible permutation it may encounter. As a result, developers are turning to synthetic data produced by existing AIs to generate the massive volume of high quality, unbiased, cheap data needed to train. (see Figure)[M] However, a recent vulnerability discovered by HiddenLayer, a leading cyber security group, determined that it is possible to send malicious pull requests with attacker-controlled data from the Hugging Face service to any repository on the platform, as well as hijack any models that are submitted through the conversion service.[M] This example illustrates how a nefarious actor can use AI generated synthetic data used to support an AI ecosystem to embed hidden commands that only trigger under certain circumstances. Essentially, the actor hides triggers inside AI generated synthetic data which can remain dormant within the system indefinitely and activated by seemingly innocuous means.

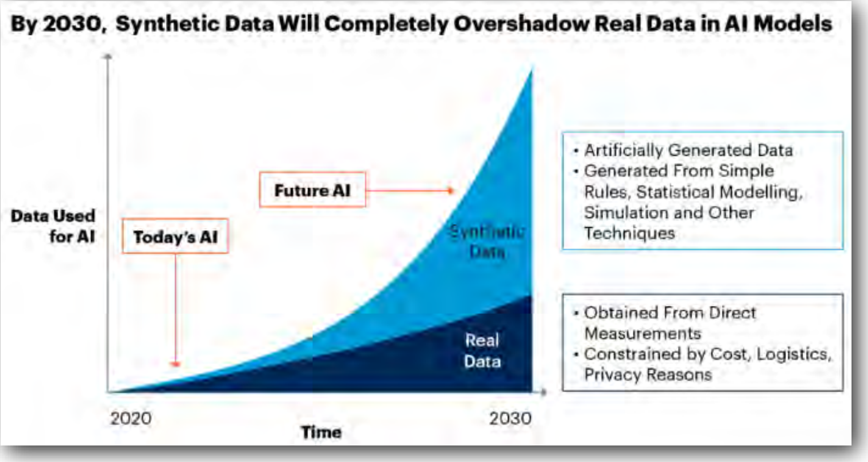


Figure 1 Use of Synthetic Data to Rise: https://www.shaped.ai/blog/how-synthetic-data-is-used-to-train-machine-learning-models. Source: Shaped AI

The model of data tampering described above relies on an activation function in the hidden layers of a neural network. (see figure 2). While not yet prevalent, the Hugging Face vulnerability demonstrates that it is possible to introduce an activation function in a little used hidden layer that will only trigger under specific conditions.[M] Specifically, hijack models submitted by users and result in data supply chain attacks, converting key data elements. In the Hugging

Face model, an AI was employed to convert data from repository to another. The AI performed its function however it hijacked the model that was supposed to be converted allowing the threat actor to make changes to any Hugging Face repository, claiming to be the conversion bot.[H] the effect was that the data was compromised. Subsequent AIs did not recognize that the data was corrupted since it was coming from a known source.

There are any number of circumstances where a hidden trigger would compromise a network. Conceivably combining the ability to hide triggers in data models with AI Generated Content results in the potential for an AI to impersonate an authorized users in the network bypassing the human defense layer. Once a trigger activates, another AI hidden in the data generates a digital facsimile of a legitimate user and employs their credentials to perform a routine function such as send or receive an email from an external source.[L] The external user might then use this access to introduce executables into the network.

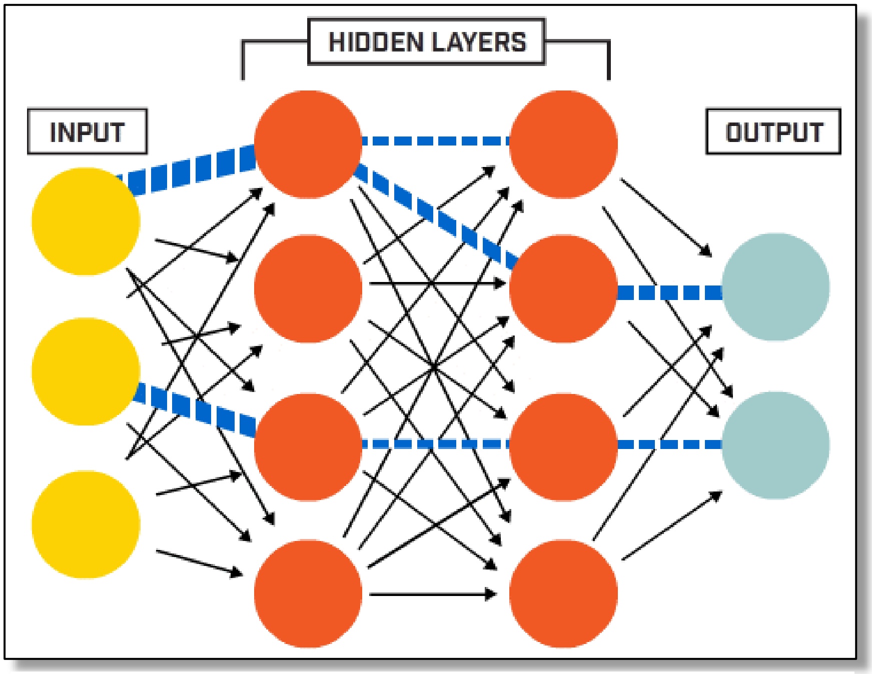


Figure 2 Introduction to Neural Networks: https://towardsai.net/p/machine-learning/introduction-to-neural-networks-and-their-key-elements-part-c-activation-functions-layers-ea8c915a9d9

This use case illustrates how AIs can encode errors into data models despite efforts to ensure the efficacy of synthetic data. While AIs are still an emerging technology, development is prolific. It is highly likely that state and non-state cyber actors will identify gaps in the Data Supply Chain where erroneous or malicious triggers can be inserted in the hidden layers and lying dormant indefinitely.[M] It is also likely that this method will be used in the near term (two to five years) to gain access to networks for later exploitation. Given the prolific nature of AI development in the near term, the best defense will be to ensure the security of the Data Supply Chain.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Given the recency of this topic, there were relatively few sources to support the final forecast. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information. It is probable that awareness of this known vulnerability will impact future cyber security conditions.

Author: COL Robert M. Richardson

Misinformation Very Likely to Lead to Mistrust in Institutions and Increase Polarization Through 2033

Executive Summary

Online misinformation is very likely (80-95%) to lead American citizens to mistrust governmental institutions and media outlets as well as drive toward continued polarization through 2033. Recent studies show that misinformation in social media will continue to increase the divide between political parties. Studies also show that the overall trust of the media is in decline, despite a slight uptick from 2022-2023, studies reveal Americans are turning away from traditional media outlets and leaning more toward social media for news.

Discussion

A combined Rutgers University and Northeastern University research study concluded that fake news, (information which is intentionally false or deliberately misleading,)[H] led some people to have higher levels of trust in governmental institutions.[H] The study found that if the fake news was supportive of one's political party, it led to higher trust. If the fake news was supportive of opposing views, it led to distrust. Figure 1 shows the findings of the study in which the trust in political institutions increased most for the most conservative adults who were exposed to fake news, while also showing a decrease in trust for the same individuals who were not exposed to fake news.[H] Katherine Ognyanova, assistant professor of communication at Rutgers University, indicated that this mistrust needs attention from multiple stakeholders.[H] Ognyanova said "Platforms should work hand in hand with media and users to implement solutions that increase the social costs of spreading fake stories. Regulators can help increase the transparency that is required in the process."[H] President Biden signed Executive Order 14110 in October 2023 which requires artificial intelligence generated content or synthetic data to be watermarked or tagged.[H] However, congress will still have to come to agreement to pass worthwhile legislation to enforce compliance.

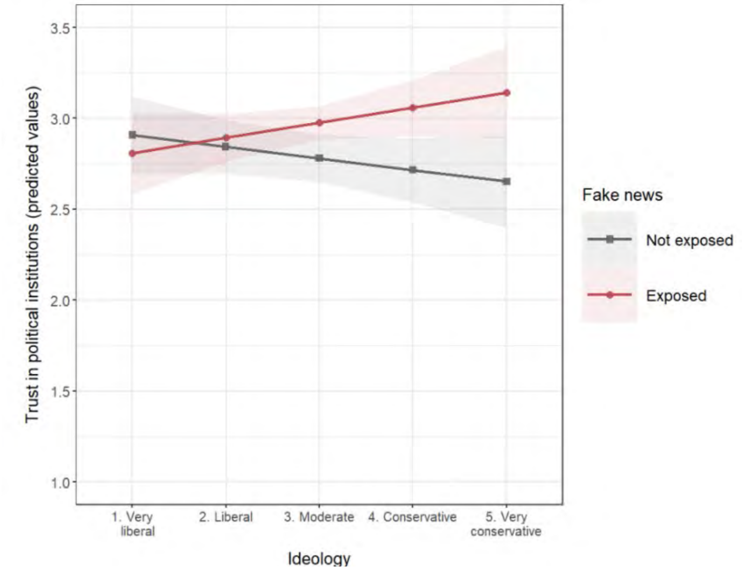


Figure 1 Rutgers University led study shows the relationship between adults who were exposed and not exposed to fake news. https://misinforeview.hks.harvard.edu/wp-content/uploads/2020/06/Misinformation-in-action-Ognyanova-et-al-2020.pdf

The Rutgers led study also discovered that online misinformation, (information which is false or inaccurate,)[H] lowered people's trust in mainstream media despite political affiliation.[H] This is corroborated by The 2022 Reuters Institute Digital News Report which showed a decline in trust of the media.[H] This report concluded that only 26% of Americans have general trust in the media.[H] As the distrust in mainstream media declines, consumers are looking more toward social media for news. In 2023, Reuters reported that consumers are more interested in social media influencers than journalists, and that young adults, aged 18-24, are making TikTok the most used social media application.[M] The 2023 report further displayed a lack of interest in traditional media to be down over 10% since 2017.[M] The United States did however see a slight increase in trust of the media in 2023 to 32%.[M] A Gallop poll from 2023 also shows American trust in the media at 32% and recognizes it as the lowest seen since 2016.[M] This poll showed a decline in the percentage of people who had a great or fair amount of trust, a slight increase in people who did not have much trust, and an increase in the percentage of people who had no trust in the media, see figure 2.[M]

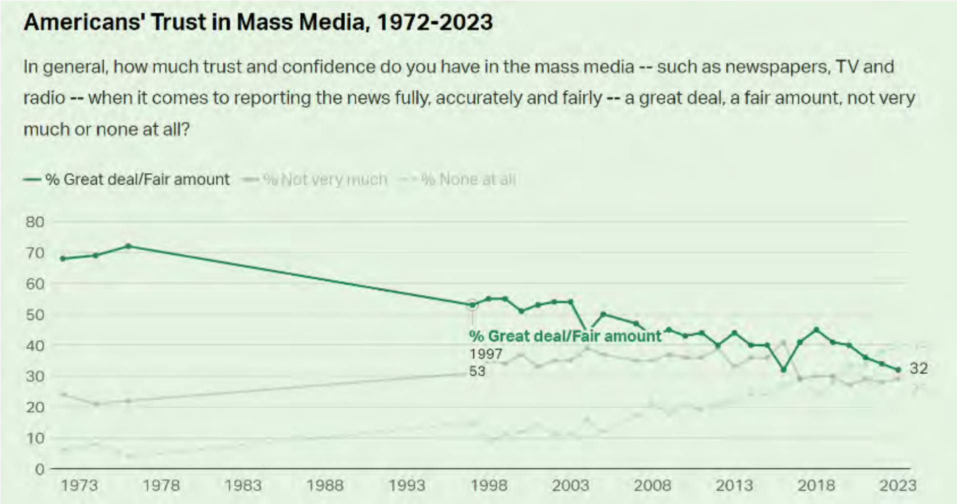


Figure 2 reveals the results of a 2023 Gallop poll showing the decline in media trust. *https://news.gallup.com/poll/512861/media-confidence-matches-2016-record-low.aspx*

A third impact from online misinformation is the increase in polarization amongst Americans. The dangers of polarization include eroding political stability and declining democratic discourse.[M] This leads to an inability of government or media to gain a consensus among citizens on major issues such as pandemic response, climate change, or domestic terrorism.[M] To complicate matters further, studies show that attempts to censor hate speech or misinformation on social media leads to further polarization.[M] It is unlikely (20-45%) that this polarization will reverse course before 2033. A study of political polarization concluded that people have a tendency to avoid information which counters their cognitive biases.[H] Even when fact-checkers are used and information is retracted, people still have a tendency to hold on to partisan beliefs and ideology.[H]

Analytic Confidence

The analytic confidence for this report is *moderate.* Sources were found to be reliable and the information was generally corroborated. The author is not a subject matter expert but looked to scholarly and peer reviewed material where available. Adequate time was available to prepare this report but the analyst worked alone without structured method. Given the time frame of this estimate, it is slightly sensitive to change due to new information such as legal legislation preventing online misinformation.

Author: LTC Charles Moss

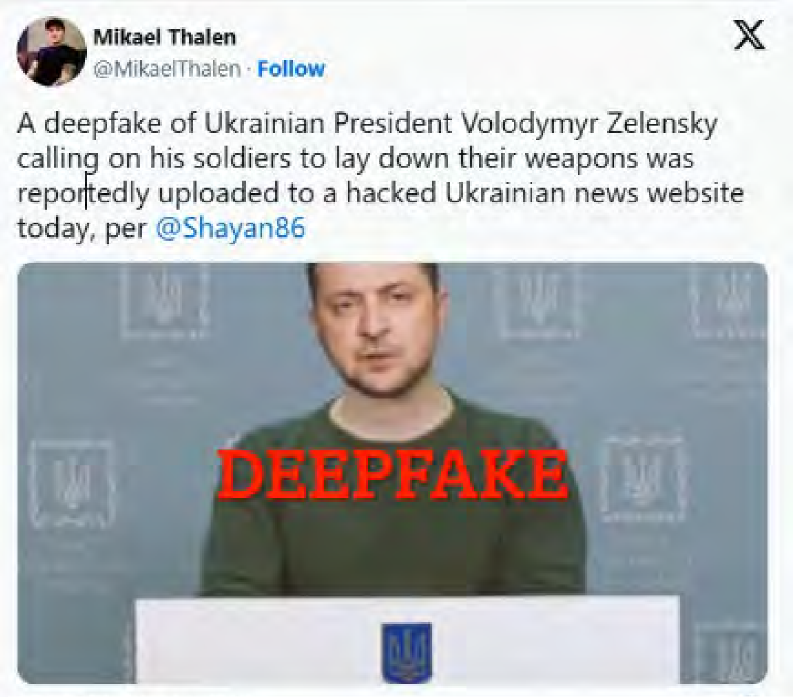
AI Tags Likely to Play a Crucial Role in Defeating Disinformation by 2033

Executive Summary

By 2030, emergent AI techniques, such as automated media tagging and enhanced watermarking, will likely (55-80%) play a crucial role in limiting disinformation. Specifically, these technologies can embed verifiable signals and metadata on the origin and creation methods in all online posts. Social platforms can then easily flag suspicious content based on these machine-readable indicators. Although privacy and accountability considerations around data access and systems governance still require resolution, it is likely that the commercial nature of AI tagging will drive the adoption cycle, along with the scalability of AI-enabled authentication offers more potential than current content-moderation approaches in the fight against viral disinformation.

Discussion

The spread of disinformation and false or misleading information presented as facts has become a crisis in the modern digital era. Social media platforms and online forums allow disinformation to spread rapidly to a wide audience, often with serious real-world consequences including eroded public trust and increased polarization.[H] Early in the Russia-Ukraine war a fake and heavily manipulated video depicting Ukrainian President Volodymyr Zelensky (See Figure 1) circulated on social media. This showed a rendering of the Ukrainian president appearing to tell his soldiers to lay down their arms and surrender the fight against Russia.[M] Ultimately, the unconvincing fake of President Zelensky was ridiculed by many Ukrainians as they quickly pointed out that Zelensky's accent was off and that his head and voice did not appear authentic upon close inspection,[M] even though this video was crude,



Mikael Thalen
@MikaelThalen · Follow

A deepfake of Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons was reportedly uploaded to a hacked Ukrainian news website today, per @Shayan86

Figure 1 Russian deepfake. Chick on picture or go to: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia Source: NPR.com

that would not be the case in the near future.

The spread of disinformation and false or misleading information presented as facts has become a crisis in the modern digital era. Social media platforms and online forums allow disinformation to spread rapidly to a wide audience, often with serious real-world consequences including eroded public trust and increased polarization.[H]

According to a recent study by the University of Bergen, human-led fact-checking alone cannot keep pace with the speed and scale of disinformation circulating today.[M] Unlike human content moderators, who can realistically review a limited number of posts per day, AI systems can instantly analyze millions of pieces of content to detect disinformation signals and trends.[H] Two technologies leading this charge are AI content tagging and enhanced watermarking protocols.[H] AI tags are invisible metadata embedded in media like news articles, social posts, or videos, encoding attributes about the origin, geo-location, creation method and more using machine learning models.[H] Watermarking similarly hides identifying codes now frequently generated by AI within the media itself rather than external tags, providing backup signals for verifying authenticity and provenance.[H]

Used in tandem, AI tagging and next-generation watermarking present a formidable solution against disinformation.[H] All media released online would carry machine-readable signals to instantly flag AI content moderation layers on major platforms, throttling the spread of inauthentic items.[H] disinformation pushers would also lose anonymity previously provided by social media, as cutting-edge watermarking traces content directly back to the source account or devices.

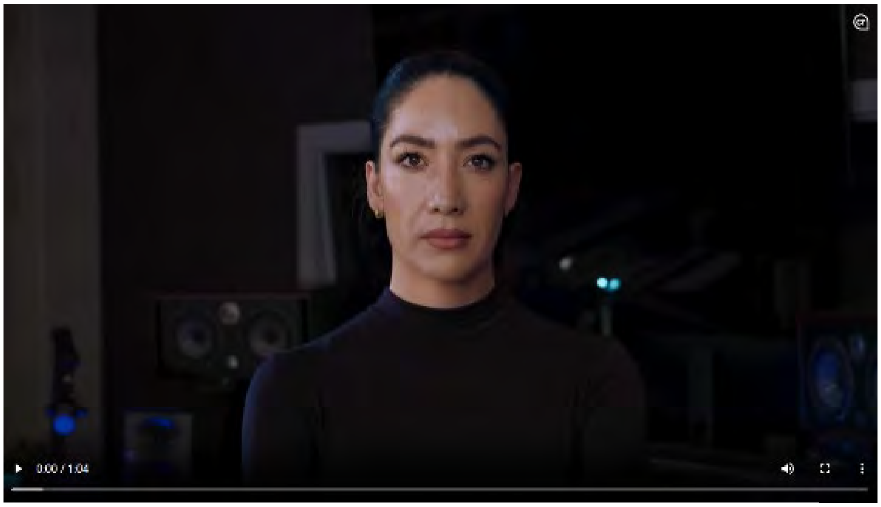


Figure 2 Example of a deepfake video with a label created by Truepic. Click on picture of go to: https://truepic.com/revel/ Source: Truepic.com

The overwhelmingly commercial nature of AI tagging makes it likely that multiple companies will be able to take advantage of advances in the technology. Projects such as C2PA named for the group that created it, the Coalition for Content Provenance and

Authenticity is a set of new technical standards and freely available code that securely labels content with information clarifying where it came from.[H] *Truepic* and acclaimed production studio, *Revel.ai*, have partnered to produce the world's first authenticated deepfake video using transparent engineering. Click the "cr" icon in the righthand corner of this video (See Figure 2) to see a version of AI-tagging.[H] The labeling is based on the latest version of a standard from the Content Authenticity Initiative led by *Adobe* and others designed to show how images and video were produced.[M] Adobe is also using similar tools to identify content created using its new Firefly generative AI tools.[M]

While the *Truepic* and *Revel* partnership provides an intriguing proof of concept for AI-authenticated media, a few other promising initiatives lend further credibility to the potential impact of technologies, such as AI tagging and watermarking. *Sentinel* headquartered in Estonia recently unveiled an AI-based protection platform its developing that serves as an authentication standard for identifying deepfakes and manipulated media, with machine-readable metadata baked into images, audio, and video files.[H] *Microsoft* is developing an Azure Content Moderator service that leverages AI to automatically apply tags and labels to identify inappropriate or offensive content across online platforms.[H] However, more research is still needed, and major technology players demonstrate consistent progress. These additional examples of AI tagging solidify an optimistic outlook for AI-enabled disinformation solutions over the next five to ten years.

Analytical Confidence

Confidence level is *moderate* that industry will adopt a tagging standard within the next 2-6 years. This assessment is based on clear evidence regarding the direction of media disinformation and misinformation technology developments. Regulatory inability to constrain growth in the near term solidifies this outlook. However, some uncertainty exists regarding potential future actions by social media platforms if issues escalate further.

Author: CDR Robert V. Liberato

Foreign Intelligence Services Will Almost Certainly Use Artificial Intelligence to Target and Recruit Susceptible Military Members by 2028

Executive Summary

Foreign intelligence services will almost certainly (95-99%) use adversarial artificial intelligence (AAI) to target and recruit susceptible United States military members by 2028. Machine learning in large language models (LLM) chatbots can manipulate the psycho-cognitive elements of human conversational interaction. Foreign intelligence services are already using AAI to target and manipulate humans for intelligence collection and exploitation. Though United States military members conduct cyber and Internet security training, an AAI chatbot designed to manipulate elements of the psycho-cognitive domain makes it almost certain a military member will be, wittingly or unwittingly, exploited using AI.

Discussion

AAI conversational chatbots effectively convince people to provide information they otherwise would not divulge. The manipulation of the human psycho-cognitive domain, their "epistemic agency," affects a person's control over their beliefs and perceptions of truth.[H] Conversational AI, through tools such as Google's AI conversational tool Language Model for Dialogue Applications (LaMDA)[H] and Open AI's ChatGPT, has accelerated to the point of conducting realistic human conversations without the human agent detecting an AI interface.[H, H] Malicious use of conversational chatbots would target human interactions to create a realistic conversation. In Figure 1, an AAI conversational chatbot (AI Agent) engages with the human user in real time. The chatbot detects the human user's responses, emotions, etc., through verbal and emotional reactions. The AAI chatbot adjusts a response in real-time, which increases the human user's perception they are speaking with another human, affecting the human user's ability on the psycho-cognitive level to discern the truth.[H] Through this function, the human user loses their "epistemic agency," which is their ability to ascertain or attribute truth.[H]

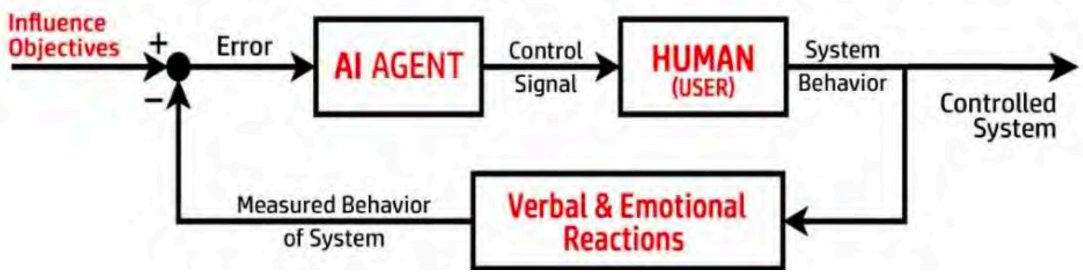


Figure 1: Information Feedback Loop for AAI Conversational Chatbot. Click on the photo to review the report or go to https://arxiv.org/pdf/2306.11748.pdf. Source: Arviv.

An AAI conversational chatbot, using this method, attacks a human's psycho-cognitive domain by gaining trust, reinforcing the human's viewpoint, and interacting on a social level.[H],[H] Cambridge psychologist Professor Sander van der Linden identifies "fluency" as one of the methods misinformation uses to target the psycho-cognitive element.[H] Fluency is when a human sees the same information repeatedly; over time, the amount of information starts to anchor as truth.[H] Researchers at Notre Dame affirmed the ability of LLM chatbots to fool human users. In a controlled experiment, the researchers found that even though the participants knew that chatbots were a part of the experiment, they correctly identified chatbots only 42 percent of the time.[H] The researchers also found that the type of persona used by the chatbot influenced participant detection. In application, this means that a foreign intelligence service can employ different personas that would very likely (80-95%) evade detection as chatbots. [H]

It is almost certain that within ten years, foreign intelligence services will predominantly rely on AAI conversational chatbots to target and exploit military members for HUMINT recruitment. Foreign intelligence services are already leveraging LLMs for collection and targeting. In a February 2024 report, Microsoft identified that Russia, China, Iran, and North Korea have all used LLMs to target personnel or create misinformation networks.[H] Microsoft assessed that this represents LLMs being used as another intelligence-gathering platform by foreign intelligence services.[H] The next step would be to combine this LLM with a chatbot persona to target specific individuals or groups. An AAI chatbot would not necessarily need to be successful for every interaction and would almost certainly target many individuals. Conceptually, the chatbot

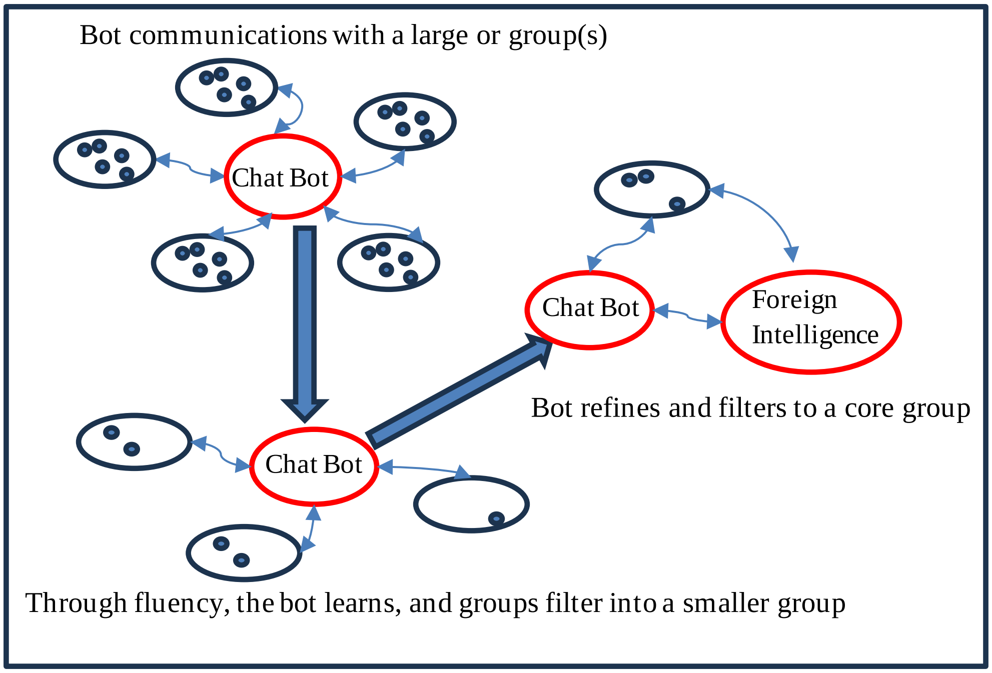

Bot communications with a large or group(s)
Chat Bot
Chat Bot
Foreign Intelligence
Bot refines and filters to a core group
Chat Bot
Through fluency, the bot learns, and groups filter into a smaller group

Figure 2: AAI conversational chatbot assessment process.

would initially target a large group (See Figure 2). As human users interfaced with the chatbot and the chatbot learned, it would continue to engage with human users until a suitable group or individual engaged consistently enough to trust the chatbot. From this point, the “relationship” would progress as a routine HUMINT source.[H] The successful recruitment of Mariam Thompson, a Department of Defense Special Operations Command linguist, gives insight into this process. A real human persona engaged with Thompson exclusively over the Internet through video and chat. Over time, the persona recruited Thompson to provide sensitive information to Hezbollah, an action Thompson would otherwise not have taken on her own.[H] A foreign intelligence service would almost certainly be capable of creating and deploying an AAI chatbot that would use this system to target United States military members.

Despite cyber training and familiarity with disinformation, it is almost certain a United States military member will be the victim of an approach by a foreign intelligence service using AAI conversational chatbots. Current cyber training relies on rules-based safeguards and awareness. It does not prepare users for a psycho-cognitive threat.[H] Because an AAI conversational chatbot targets a person’s psycho-cognitive element and induces trust through fluency, cyber training safeguards would be ineffective in preventing exploitation. Cambridge University Professor of Psychology Sander van der Linden posits that “inoculation,” small doses of misinformation to train the human user on what wrong looks like, is the best defense at the cognitive level.[H] The National Institute of Science and Technology also recommends adopting a “zero trust” mindset to create mental armor.[H] Without updated training that explicitly addresses an AAI’s ability to target and exploit at the psycho-cognitive level, it is almost certain that a foreign intelligence service will recruit a military member through a conversational chatbot within five years.

Analytic Confidence

The analytic confidence for this estimate is *high*. Sources were reliable and tended to corroborate one another. Perplexity was used, and ideas from the results were utilized in further research. Perplexity suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Given the lengthy time frame of the estimate, this report is sensitive to change due to new information, such as advances in LLMs and the propensity for foreign intelligence services to accelerate experimentation with LLMs and machine learning.

Author: Mr. Tom M. Jackson

ERGO SUM
MACHINA

Additional Findings

Adversaries Highly Likely to use Artificial Intelligence Generated Content to Gain Access to Secure Networks Despite Emerging Policy and Legislation by 2028

Executive Summary

The proliferation of Artificial Intelligence-Generated Content (AIGC) will almost certainly (95-99%) enable adversaries to revolutionize malicious cyber operations to infiltrate networks, augment cyber-attacks, and evolve cyber espionage methodologies despite AI-enabled cyber security. Despite emerging legislation and policy, AIGC will enable cyber actors to conduct refined attacks at scale to overwhelm or bypass existing cyber protections and likely (55-80%) poison synthetic data pools used to train AIs to infiltrate networks at the machine level bypassing human security.

Discussion

Artificial Intelligence (AI) has seen significant growth and transformation in recent years and will likely (55-80%) grow exponentially through 2030. According to a 2021 report from Stanford Institute for Human-Centered Artificial Intelligence, AI has made substantial strides in large data review on specific tasks. [H] By 2024, 80 percent of corporations plan on incorporating AI within two years.[M] See Figure 1 for more information on corporate adoption of AI. According to market estimates, the future of AI is projected to expand from $150.2 billion in 2023 to more than $1,345 billion by 2030.[M] Unfortunately criminal and adversarial state cyber operations are poised to take advantage of the growth in AI to augment cyber activities.

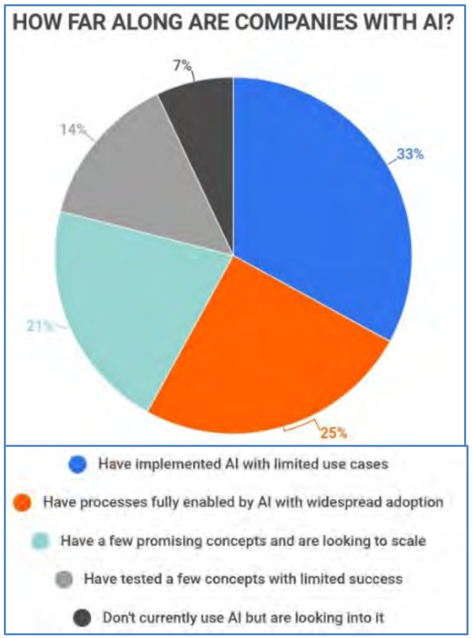


Figure 1 Corporate AI Adoption Source: https://www.zippia.com/advice/artificial-intelligence-statistics/#AI_Market_Trends_and_Predictions

State and non-state cyber actors have already begun to incorporate AI tools and AI Generated Content into traditional malicious cyber-activity such as Phishing, Brute Force Attacks on critical infrastructure, and intellectual property theft. [H,H,M,] Similarly, deepfakes, false news, and social media reporting are becoming more prevalent with both states and cybercriminals incorporating AI tools and techniques into malicious cyber activities with prospects for significant growth over the next five years.[H,H,M,M] According to most leading cyber security specialists, the assessment is that despite the emergence of policy, legislation, and international norms, cyber protections are unlikely to keep pace with hostile cyber activity.[M,]

[M], [L] As a result of the lag in policy and international norms, current protections are entirely in the technical domain, and cybercriminals are innovating faster than protections can adapt.[H], [M]

Cybercrime is already lucrative and difficult to prosecute despite the relative ease of attribution and is very likely to grow over next few years.[H] The United States alone suffered $10.3 billion in losses in 2022. At the current growth rate, losses will exceed ~$92 billion by 2028.[H] It is also difficult prosecute cybercriminals given that most cybercriminals generally reside outside the borders of the state they are active in to avoid arrest. Further, a declining number of interdictions combined with increasing financial losses mean that cyber criminals are becoming more effective.[H] Incorporating AIGC into Social Engineering attacks, false Tech Support, Extortion, Non-Payment / Non-Delivery, Personal Data Breach, and Phishing attacks will likely only increase the speed and volume of attacks exponentially.[9], [H] Cybercriminals will likely be drawn to AIGC enabled techniques to enable moving operations to scale.

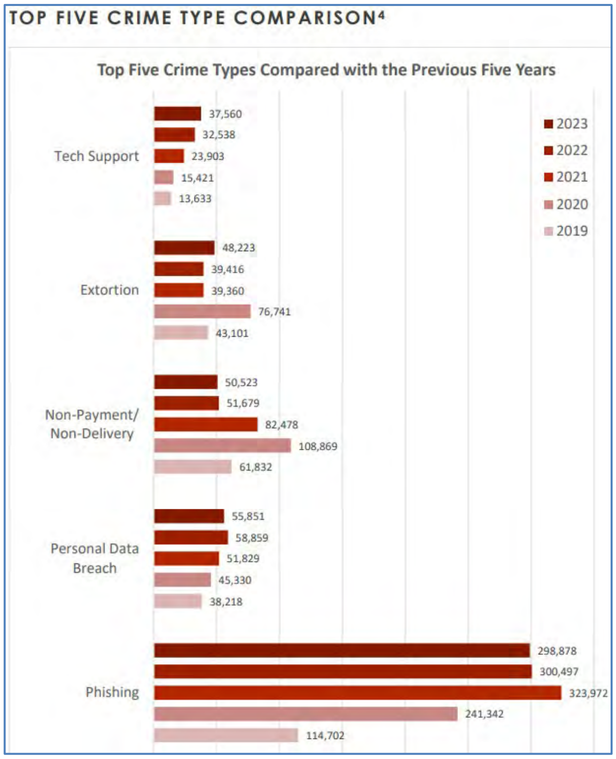


Figure 2: FBI 5 Yr Cyber Crime Report. Source: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf

States that harbor cybercriminals will almost certainly exploit emerging techniques and innovate novel criminal uses of existing tools. However, sophisticated states have other advantages and will very likely use AIGC in novel approaches to mature their cyber espionage operations. China is very likely (80-95%) to enhance its cyber operations with AIGC, increasing the sophistication and volume of attacks by 2029 with a focus on gaining and maintaining access to networks.[H] China's robust cyber capability, combined with AI, poses a formidable threat. Despite international efforts to establish norms and controls, China's capability will likely outpace cyber protections over the next five years.[H] Taking another approach, Russia is restructuring its cyber capability around a new National AI Center with the stated purpose of weaponizing AI. As a result, Russian cyber operations are

[9] 2023 FBI Internet Crime Report lists the top five most prevalent categories of cybercrime activity as Tech Support (impersonating IT support to gain access to user accounts), Extortion, Non-Payment / Non-Delivery (data ransom), Personal Data Breach, and Phishing.

also anticipated to evolve significantly, shifting from brute force attacks to more refined cyber espionage by 2030.[M]

The current applications as identified by Microsoft and other cybersecurity firms have detected the use of generative AI by U.S. adversaries, including Iran, North Korea, Russia, and China, in offensive cyber operations.[H] These operations aim to breach networks and conduct influence operations, with generative AI enhancing the sophistication of deepfakes, voice cloning, and social engineering.[H,M,M] AI-generated attacks leverage artificial intelligence and natural language processing to deceive and compromise individuals, organizations, and systems.[H,M] These attacks are becoming increasingly sophisticated, mimicking the language and style of legitimate communications to bypass traditional security measures.

While it is almost certain that AI will enable malicious cyber activity over the next five years, it is also nearly certain (95-99%) future adversaries will leverage AIGC to enable sabotage of critical vulnerable systems.[H,M,M] There are subsets of AIGC that will become increasingly vulnerable to malicious manipulation that will be difficult to detect and will bypass many existing cyber security measures.[M] Generally, AIGC includes synthetic media, which encompasses any AI-generated media, including images, videos, texts, and audio however, synthetic data is another form of generated content produced by an AI and is representative of real data.[H,H] Exploiting synthetic data and AI managed protocols will very likely (80-95%) be an area of future growth among sophisticated adversaries. The utilization of AI-generated synthetic data for training AI systems is projected to significantly increase the risk to network security between 2025 and 2030.[M] The poor efficacy of the data supply chain, despite cyber protections, is a primary concern. Vulnerabilities, such as those discovered in the Hugging Face service, illustrate how attackers can embed hidden commands within synthetic data, potentially compromising networks indefinitely.[M,H] As a result, it will become increasingly vital to protect the Data Supply Chain down stream in order to protect the integrity of AI's and AI models that drive

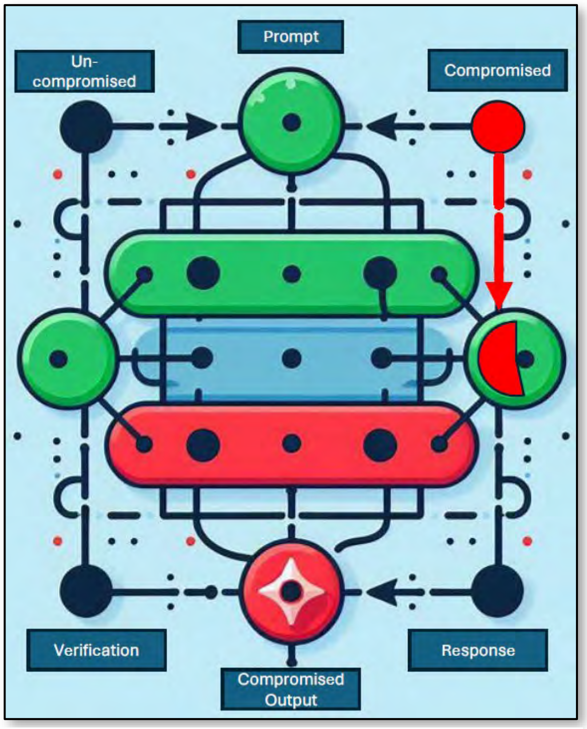


Figure 1 Simple Depiction of Compromised Node in Neural Network. Source: Generated by DALL-E

future systems. Exploiting synthetic data or manipulating the process layers of neural networks in future AI enabled networks will likely be constrained to adversaries with sophisticated cyber capabilities. The primary delimiting factor for use of AIGC to tamper with synthetic data or manipulate neural networks is ready access to supercomputing power.

The inclusion of Generative AI into cyber operations will very likely drive a change in cyber espionage targeting methodologies, posing an increased risk to corporate and industry leaders despite workplace cyber protections.[M] The leveraging of AI and AIGC will likely bypass most existing institutional security through sophisticated targeting approaches.[M] The malicious use of AIGC represents a rapidly evolving threat landscape. Adversaries are likely to exploit AIGC for infiltrating networks, augmenting state-level cyber operations, and evolving cyber espionage tactics.[H,M,M] The proliferation of AI technologies, coupled with the sophistication of attacks, underscores the urgent need for robust data supply chain security and international cooperation to mitigate these emerging threats.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information.

Author: COL Robert M. Richardson

Revolutionizing Defense: U.S. Government Very Likely to Use Machine Learning by 2033 for Combating AI-Generated Threats, Despite Lingering Vulnerabilities

Executive Summary

By 2033 it is very likely (80-95%) the U.S. government will harness advancements in machine learning for detecting and attributing adversarial use of artificial intelligence generated content (AIGC). The U.S. military will very likely (80-95%) execute a number of countermeasures due to almost certain (95-99%) adversarial use of AIGC in the information sphere during conflict. These actions include training in enhanced cyber and AI literacy, as well as investing in machine learning algorithms for detection and attribution. Despite U.S. countermeasures for adversarial AIGC, there will remain high-risk target areas, specifically the SCADA systems that support civil infrastructure.

Discussion

In October 2023, the Office of the Director of National Intelligence pitched a “proposal’s day” briefing to industry, both commercial and private sectors, titled “Securing our Underlying Resource in Cyber Environments (SoURCE).[M] The program is seeking novel technologies that will make determinations of the potential attackers, based on coding styles, and then measure the similarity between files to provide suspected origins (country, groups, individuals, etc.).[M] It is very likely (80-95%) after testing and perfecting algorithms and software in the U.S. National laboratories, as a result of continued collaboration with industry, that the U.S. government will utilize machine learning, a type of AI[M], to pinpoint cyber attribution.

In recent years psychological operations have been synonymous with military information support operations (MISO), using technology to influence perceptions, shape behaviors, and guide decision-making processes among target audiences.[M] The rise of AIGC has directly impacted digital psychological operations. In 2024, the possibility for AIGC to influence and spread misinformation and disinformation is likely (55-80%) the highest it has been since its inception due to the high number of global

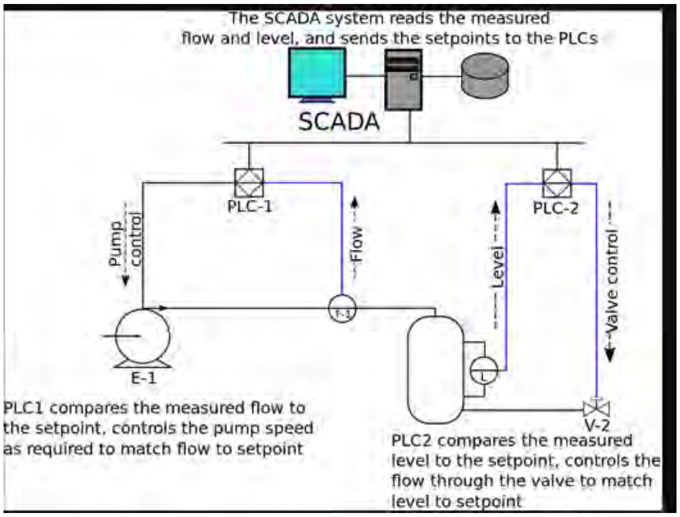

The SCADA system reads the measured flow and level, and sends the setpoints to the PLCs
SCADA
PLC-1
PLC-2
Pump control
Flow
Level
Valve control
F-1
E-1
L
V-2
PLC1 compares the measured flow to the setpoint, controls the pump speed as required to match flow to setpoint
PLC2 compares the measured level to the setpoint, controls the flow through the valve to match level to setpoint

Figure 6: Example of a SCADA schematic, referenced in the paper "Cybersecurity and the U.S. Energy Grid," by Nikhil Partasarathy submitted as coursework (Stanford University, Fall, 2016).

elections.[M] Countries who may have been just beginning to use technology years ago can now do more operations, do them faster, and produce more catastrophic effects, with even less attribution.

The almost certain use of adversarial use of AIGC in the information domain will very likely cause the U.S. military to enhance professional military education (PME) and training to incorporate AIGC literacy and training. Additionally, the use of detection software, which can check the veracity of open source data and possibly identify videos and images as fake, will likely be employed as it can mitigate risk.[H]

Understanding how adversaries might use AIGC to target the U.S., it is almost certain that the U.S. intelligence community (IC) will be acutely focused on protecting data collection from any potential AIGC data poisoning. Checking the veracity of data, given the deepfake threats, is already common practice in IC agencies.[M] Moreover, government agencies have been aware of this threat since 2021,[M] so it is likely that detection tools or safeguarding measures have already been put in place.

Despite the many countermeasures, there are still vulnerabilities which adversarial AIGC will likely target, such as the U.S. Supervisory Control and Data Acquisition (SCADA) systems. SCADA systems are the "brains" of an industrial control system and have transitioned from standalone entities to now being connected to the internet for more efficient communication of data.[H] The transition to a highly interconnected network has made SCADA more vulnerable to various cyber-attacks, given that security approaches provided by IT-based systems are not efficient enough to detect the risks and threats.[H]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change. Intelligence reporting at a higher classification level could affect this forecast, for instance if agencies within the intelligence community have already begun working on countermeasures or employing detection software against AIGC.

Author: LTC Katherine M. Ogletree

Psycho-Cognitive Domain Unlikely to be Significantly Affected by Online Content Through 2033

Executive Summary

Adversarial use of AIGC in media is unlikely (20-45%) to significantly affect the psycho-cognitive domain. This is due to the difficulties which occur when countering cognitive bias and the involuntary human response which occurs in the brain when given new or conflicting information. Further, theories of truth have evolved away from being fact based and are correlated to emotion, making it more difficult to affect what someone holds true. Adversaries have made attempts to influence elections by attacking the psycho-cognitive domain, but their efforts have seemingly failed. However, prolonged exposure to online misinformation and disinformation is very likely (80-95%) to have the effect of reducing trust in government institutions and mass media.

Discussion

A Harvard Business Review study from 2020 revealed that cognitive bias and personal history are extremely difficult to overcome, regardless of soundness of argument or presentation style.[H] This study was corroborated by a Berkley study which found that humans have a strong desire to maintain pre-existing beliefs and convictions.[M] The Berkley study also concluded that when a human is presented with facts which counter their cognitive biases, they have an automatic tendency to strengthen their previous beliefs.[M] This occurs due to a chemical reaction in the brain, specifically in the amygdala. The amygdala interprets changes in information as a threat and releases hormones to help the human body prepare for fear, fight, or flight.[M]

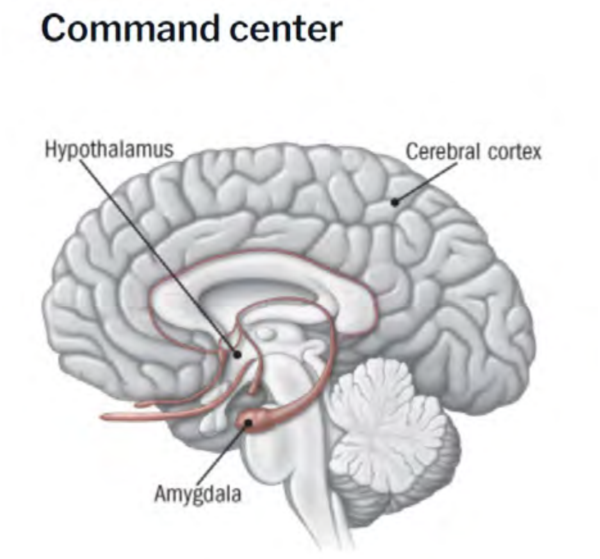


Figure 7 Harvard Medical School depiction of how the brain responds to stress signals. https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response

A 2020 Harvard Medical School study corroborates this. The amygdala interprets new or conflicting information as dangerous and sends a distress signal to the hypothalamus.[H] Meanwhile the hypothalamus acts as a command center and communicates with the rest of the nervous system to give a person the ability to fight or flee, see figure 1.[H] The human body actively works against receiving information which is counter to pre-conceived notions. It is therefore unlikely the adversarial use of artificial intelligence generated content (AIGC) would result in an adult changing their mind or behavior from misinformation or disinformation.

The pre-conceived notions and cognitive biases people believe do not necessarily need to be factual for the believer to hold them true. The pragmatic theory of truth is summed up by concluding that something is true if it is useful and untrue if it is unusable.[H] This theory helps explain why a person can believe something which has little or no basis in reality, they simply find it useful in their life. This also helps corroborate the Harvard and Berkley studies concerning the difficulties in changing someone's mind as well as their fight or flight reaction to information which challenges their heuristics. According to the National Library of Medicine, a major determinate of a media consumer believing what they see or read is confirmation bias.[H] This pushes people to seek and interpret information which is in accordance with their existing beliefs and expectations.[H] This theory reinforces that it is unlikely an adult would have significant change in their perception of truth regardless of media content.

The Chinese Communist Party (CCP) engaged in an online disinformation campaign using AIGC to attempt to influence the 2024 presidential election in Taiwan.[H] The CCP used AIGC in hopes of getting their preferred candidate, Hou Yu-ih from the Kuomintang Nationalist Party or KMT, elected.[H] However the efforts from the Chinese Communist Party failed and their least preferable candidate, Lai Ching-te from the Democratic Progressive Party (DPP,) won.[H] Figure 2 shows the election results. China wanted to prevent the DPP from winning because they support Taiwan independence from the People's Republic of China. Conversely, the KMT was the preferred party because it is pro-China.[M] Despite the CCP efforts, online content failed to have such a significant impact that it steered the election in their favor.

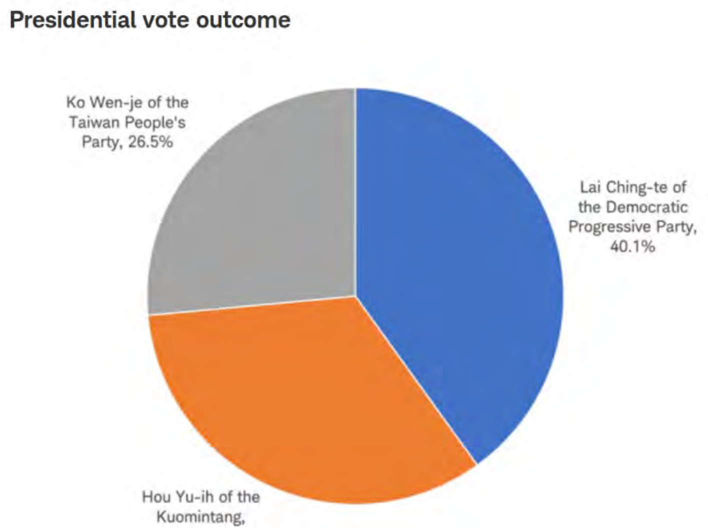


Figure 8 Results of the 2024 Taiwan presidential election. *https://www.schwab.com/learn/story/global-impact-taiwans-election*

Despite online content not significantly shaping human perception, online misinformation and disinformation will very likely lead to reduced trust in governmental institutions and mainstream media. A combined Rutgers University and Northeastern University research study concluded that fake news, (information which is intentionally false or deliberately misleading,)[H] led people to generate their biases toward the political party they were most closely aligned.[H] This created distrust amongst those whose party was not in charge and increased trust for those whose party was in charge. Trust in mass media has had a downward trend over the last 40 years.[M]

Analytic Confidence

The analytic confidence for this report is *moderate*. Sources were found to be reliable and the information was generally corroborated. The author is not a subject matter expert but looked to scholarly and peer reviewed material where available. Adequate time was available to prepare this report, but the analyst worked alone without structured method. Given the time frame of this estimate, it is sensitive to change which could be brought about by new theories or explanations of how humans accept new information or actors finding new ways to appeal to opposing biases without the individual being aware.

Author: LTC Charles Moss

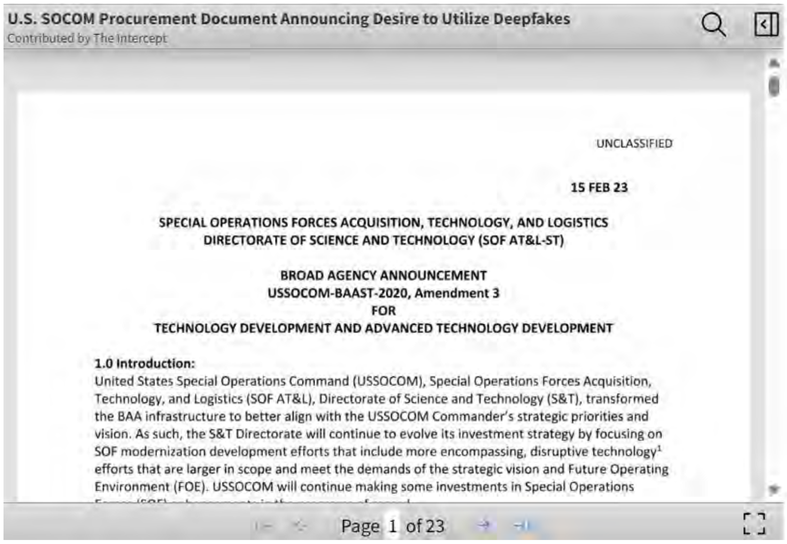
The DOD Will Likely Have a Force Ready to Combat Deceptive Practices in the Psycho-Cognitive Domain by 2033

Executive Summary

Due to the rapidly growing disinformation threat and the DOD's recognition of the severity of the threat and investment in counter-disinformation technologies. The Department of Defense (DOD) is likely (55-80%) to have a force ready to combat deceptive practices enabled by artificial intelligence (AI) by 2033. Despite the increasing sophistication of adversary AI-driven disinformation capabilities, the DOD's focused efforts and organizational momentum suggest that it will be well positioned to fight and win in the psycho-cognitive domain by 2033.

Discussion

The psycho-cognitive domain is evolving at an alarming pace due to the advent of advanced technologies and proliferation of information.[H] Over the next five to ten years generative adversarial networks (GANs) will likely enable increasingly realistic fake content generation from images and audio to video.[H] Similarly, disinformation bots powered by advanced language models will be able to engage in highly tailored, emotionally manipulative influence campaigns at scales that are not possible with human operators alone.[H] Adversaries will soon be able to leverage digital platforms, artificial intelligence, and data manipulation to disseminate disinformation, conduct psychological operations, and undermine their targets decision-making processes. The weaponization of information has become a critical component of modern conflict, blurring the lines between traditional and nontraditional forms of warfare.[H]


U.S. SOCOM Procurement Document Announcing Desire to Utilize Deepfakes

Contributed by The Intercept

UNCLASSIFIED

15 FEB 23

SPECIAL OPERATIONS FORCES ACQUISITION, TECHNOLOGY, AND LOGISTICS
DIRECTORATE OF SCIENCE AND TECHNOLOGY (SOF AT&L-ST)

BROAD AGENCY ANNOUNCEMENT
USSOCOM-BAAST-2020, Amendment 3
FOR
TECHNOLOGY DEVELOPMENT AND ADVANCED TECHNOLOGY DEVELOPMENT

1.0 Introduction:
United States Special Operations Command (USSOCOM), Special Operations Forces Acquisition, Technology, and Logistics (SOF AT&L), Directorate of Science and Technology (S&T), transformed the BAA infrastructure to better align with the USSOCOM Commander's strategic priorities and vision. As such, the S&T Directorate will continue to evolve its investment strategy by focusing on SOF modernization development efforts that include more encompassing, disruptive technology[1] efforts that are larger in scope and meet the demands of the strategic vision and Future Operating Environment (FOE). USSOCOM will continue making some investments in Special Operations

Figure 1 United States Special Operations Command desires to develop new capabilities in influence operations. Click picture for link to document. https://www.documentcloud.org/documents/23696654-us-socom-procurement-document-announcing-desire-to-utilize-deepfakes Source: The Intercept

Recognizing the importance of this threat, the DOD will likely prioritize the development of capabilities to combat deceptive practices. The 2022 National Defense Strategy explicitly acknowledges the need to "compete, deter, and win" in the information environment, highlighting the importance of building resilience against adversary influence operations.[H] This strategic emphasis sets the stage for the DOD to allocate resources and focus its efforts on developing a force ready to combat deceptive practices by 2033. This is emphasized by the United States Special Operations Commands proposal (Figure 1) for additional capabilities to utilize deepfakes and tools to counter them.[H]

To effectively combat deceptive practices, the DOD is likely to explore cutting-edge technologies that can detect, analyze, and counter adversarial tactics. Artificial intelligence and machine learning algorithms will play a crucial role in this endeavor, enabling rapid identification and analysis of deceptive content across vast datasets. Promising AI-based detection algorithms are already being developed in cooperation with the DOD, such as *Massachusetts Institute of Technology (MIT)* Lincoln Lab's Reconnaissance of Influence Operations program, to automatically identify deep fakes and disinformation narratives.[H] Enhanced digital watermarking and content provenance standards such as project C2PA named for the group that created it, the *Coalition for Content Provenance and Authenticity* offer scalable authentication solutions.[H] Additionally, the DOD has recognized the potential of these technologies and has initiated programs such as the *Defense Advanced Research Projects Agency's (DARPA)* Semantic Forensics (SemaFor) project, which aims to develop innovative semantic technologies for analyzing media.[M] These technologies include semantic detection algorithms, which will determine if multi-modal media assets have been generated or manipulated. Attribution algorithms will infer if multi-modal media originates from a particular organization or individual. Characterization algorithms will reason about whether multi-modal media was generated or manipulated for malicious purposes. These SemaFor technologies will help detect, attribute, and characterize adversary disinformation campaigns.[H] (See figure 2 for more information). Furthermore, the commercial industry is investing in the development of advanced analytics and data visualization tools that can help decision makers quickly identify patterns and anomalies in the information environment.[H] Social media platforms, such as *Facebook*, are showing interest in labeling AI-generated content, which is another positive sign.[H] Additionally, automated media tagging and

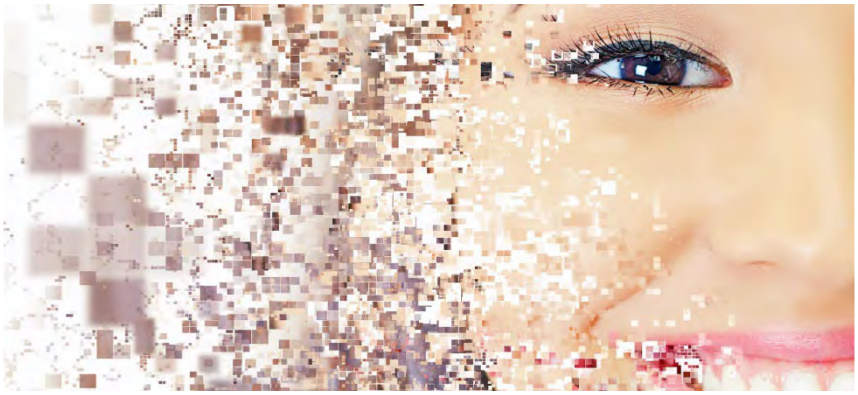


Figure 2 *DARPA* SemaFor Project. Click picture for additional information. https://youtu.be/UW0VYZzSgPY?si=fEW8lLcGzs4JOybl Source: DARPATv via Youtube.com

enhanced watermarking will play a crucial role in limiting disinformation.[H] These tools likely will enable the DOD to proactively detect and respond to deceptive practices, thereby minimizing their impact on military operations and decision-making processes.

Beyond technology, a skilled and adaptable workforce is essential to fighting deceptive practices. The DOD is likely to prioritize the recruitment, training, and retention of personnel with expertise in fields such as data science, cybersecurity, and psychological operations.[H] These individuals form the backbone of the force tasked with combating deceptive practices, bringing about a deep understanding of adversarial tactics and the ability to develop innovative countermeasures. To ensure that the force remains agile and responsive to evolving threats, DOD is likely to foster a culture of continuous learning and adaptability.[H] Regular training and education programs will be essential to keep personnel updated on the latest deceptive practices and countermeasures. The Air Force Culture and Language Center has begun to addresses this through an educational video series on its Culture Guide app focused on helping total force Airmen and the Department of Defense develop resilience to misinformation and disinformation.[H]

Combating deceptive practices is not a task the DOD can undertake alone. The global nature of this threat necessitates strong partnerships and alliances with domestic and international entities. The Department of Education is likely to promote media literacy to help the public recognize deep fakes and disinformation using lessons learned from places that have had success such as Finland’s development of a strong education program to resist misinformation recognizing the need, in California, Assembly Bill 873 was recently passed that requires the state to add media literacy curriculum for all K-12 students.[H] On the international front, and the DOD is likely to strengthen alliances with like-minded nations facing similar challenges. Collaborative efforts, such as joint training exercises, information sharing, and the development of interoperable systems, are crucial for building a collective defense against deceptive practices. These partnerships will not only enhance the DOD's capabilities, but also serve as a deterrent to adversaries seeking to exploit vulnerabilities in the information environment.

While the DOD's commitment to combating deceptive practices is evident, the rapid pace of technological advancement means that adversaries will continue to develop new and sophisticated deceptive tactics, thus requiring the DOD to maintain a constant state of vigilance and adaptability. Despite the DOD's significant investments in advanced technologies and its commitment to developing a skilled workforce, the rapidly evolving nature of AI-driven disinformation tactics may pose a constant challenge, requiring the DOD to remain vigilant and adaptable in order to stay ahead of adversaries' ever-changing strategies. Despite these challenges, the DOD will likely lead the way in developing a force capable of effectively combating deceptive practices.

Analytical Confidence

The confidence level is *moderate*, and the inherent unpredictability and rapid evolution of the AI field mean that specific predictions beyond 5 years carry uncertainty. This assessment is based on clear evidence regarding the direction of media disinformation and the development of misinformation technology. Confidence increases with additional information on DoD budgetary plans, organizational changes, and R&D priorities related to counter disinformation. More robust intelligence assessments of adversaries' AI-enabled disinformation programs would also help solidify the outlook and clarify comparative United States advantages or deficits over time.

Author: CDR Robert V. Liberato

Software Almost Certain to be the Tool Used to Detect Artificial Intelligence Generated Content through 2033

Executive Summary

Artificial intelligence detection will almost certainly (95-99%) be done by software through 2033. The leading artificial intelligence generated content detection is almost 80% accurate for text and 65% for video. Artificial intelligence is capable of being misused by users to cheat in school and manipulate the way people think for nefarious use, to include military gains. The weaponization of misinformation is not a new concept but social media makes the spread faster and has the potential to reach a wider audience. However, detection software is subject to false positives with some incidents up to 50%.

Discussion

The solution to detecting artificial intelligence generated content (AIGC) will almost certainly be AI driven software. Text AI detecting software is designed to compare perplexity and burstiness.[M] Perplexity refers to the predictability of the text and determines if it has a natural flow and makes sense whereas burstiness compares the variances in the sentence structure and length.[M] Currently, AI detectors are not perfectly accurate. A study from the International Journal of Academic Integrity showed that AI detection was less than 80% accurate.[H] This study also showed that some detection tools gave up to 50% false positive results,[H] the tool identified work as AIGC but was not. AI generated video detection is not as far developed as text detection. In 2021, the winner of the Deepfake Detection Challenge was able to decipher deepfake videos with 65.18% accuracy.[H] A Brookings Institute study revealed that post-hoc image, audio, and video detection software ranged from 3-96% accuracy and stated that detection software will be part of a layered approach to AIGC detection.[H]

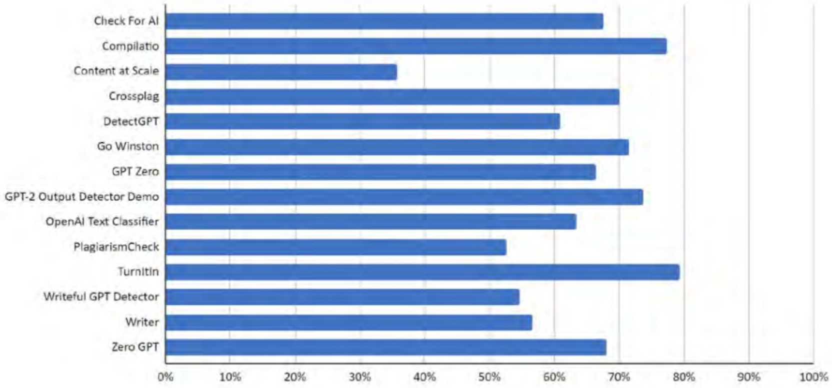


Figure 9 shows the results of a study from the International Journal of Academic Integrity. *https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00146-z*

Since AI became available to the public, users have found ways to misuse it. OpenAI's ChatGPT was openly released in November 2022.[H] Almost immediately, students ranging from elementary school through post-secondary education began using it to write their assignments for them.[H] AI progressed from text generated content and users can now have a program create AIGC in photo and video format. This evolution of AIGC allows anyone with internet access to create videos of almost anyone saying or doing anything. The purposes of using malicious deepfakes range from breaking will, discrediting, extortion, creating time, intimidation, and creating confusion.[H] In an effort to both break the will of the Ukrainians and create confusion, a deepfake video showed President Zelensky telling Ukrainians to surrender, see figure 2.[H] The video was discredited and did not achieve the desired effect but as AI technology improves, the detection will become increasing difficult.

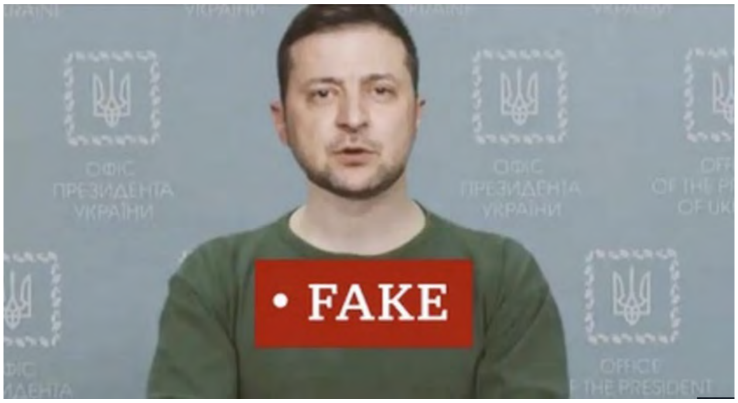


Figure 2 shows a clip from a Russian deepfake video of Ukranian President Zenensky. https://www.bbc.com/news/technology-60780142

The United States Federal Bureau of Investigation (FBI) Director Christopher Wray indicated misinformation is not a new concept but the proliferation of AIGC on social media, coupled with the United States population's growing reliance on social media creates a new escalation of weaponizing misinformation.[H] To help counter this, several United States based social media companies already have policies in place to help prevent the spread of misinformation. Meta, the parent company of Facebook and Instagram, and X, formerly Twitter, both reserve the right to remove or mark AIGC misinformation when it is found.[H] [H] Further, President Biden signed Executive Order 14110 in October 2023 which will require watermarks or labels for AIGC or synthetic data.[H]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Most sources were found to be highly reliable. The moderately reliable sources were from previously unknown sites, but the information tended to corroborate over multiple sources. Adequate time was allowed but the analyst worked alone to research and compile information without a structured method. Despite the lengthy time frame of the estimate, this report is subject to change due to pending legislation and advancements in future software which could show increased difficulty in detection.

Author: LTC Charles Moss

Generative Adversarial Networks Very Likely to Increase Threats By 2033

Executive Summary

As Artificial Intelligence (AI) capabilities become more powerful and widespread, generative adversarial networks (GANs) will very likely (80-95%) enable new varieties of attacks and threats to US forces by 2033. GANs are advancing rapidly and will reach new scalability and accessibility thresholds by 2033. This will significantly expand the threat landscape despite the limitations in predicting long-term AI developments. The scalable generation of realistic fake media provides adversaries with new avenues for influence campaigns, cyber warfare and information warfare.

Discussion

As Artificial Intelligence (AI) capabilities become more powerful and widespread, it is very likely that the growing use of AI systems will lead to an expansion of existing threats.[H] As a result, it will be worthwhile to attack targets that otherwise would not make sense from the standpoint of prioritization or cost-benefit analysis.[H] This will broaden the range of groups capable of conducting such acts, enabling more frequent attacks, new types of attacks, and expanding the set of vulnerable targets. These attacks may use AI systems to complete certain tasks more successfully than any human can apply deceptive techniques, such as mass deception, emotional manipulation, and predatory targeting. [H]

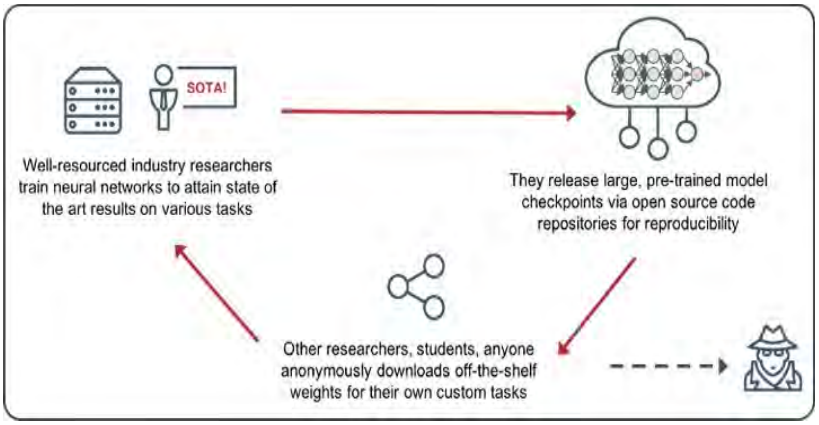

SOTA!
Well-resourced industry researchers train neural networks to attain state of the art results on various tasks
They release large, pre-trained model checkpoints via open source code repositories for reproducibility
Other researchers, students, anyone anonymously downloads off-the-shelf weights for their own custom tasks

Figure 10 The culture of machine learning research has given rise to a rich open source model ecosystem. Source: Mandiant.com

One area of (AI) seeing giant leaps forward is generative adversarial networks (GANs), AI system comprising two neural networks that work against each other to refine the realism of synthetic media.[H] GANs have shown astonishing progress for the entertainment industry in creating compelling fake images and videos that are extremely difficult for humans or AI to detect as false. These algorithms are steadily improving, ushering in an era where realistic deep fakes can be produced at a scale with minimal resources and technical skill. (See Figure 1)

Although GANs hold great promise, they also open the door for new forms of deception and manipulation. The adoption of GANs is changing the nature of security risks faced by governments and militaries. From a security perspective, several of these developments are noteworthy. For instance, the ability to recognize a target's face and navigate[H] through airspace can be applied to autonomous weapon systems. Similarly, the ability to generate synthetic images, text, and audio can be used to impersonate others online (See Figure 2)[H] or to sway public opinion by distributing generated content through social media channels.[H]

Recent real-world cases have revealed malicious use of GANs. Adversaries such as Russia deployed this technology in their war with Ukraine. For instance, in 2022, the Lieber Institute at West Point reported that Russian operatives deployed convincing videos of political figures through social media to influence policy and sow political discord.[H] The videos were believed to have been GAN-manipulated. Additionally, there have been examples of using GANs to produce nonconsensual fake intimate imagery, as well as AI-generated images and video materials that aid online harassment campaigns while protecting the identity of the perpetrators.

Generating synthetic media allows attackers to hide behind AI systems and evade attribution.[H] Highly realistic fakes are more likely (55-80%) to achieve adversarial goals such as manipulating public opinion or gaining access to sensitive systems. Producing volume output disguises, the origin point and overwhelms investigators attempting to separate the real from the fake ones.[H]

GAN systems are likely to unlock new threats that are inconceivable with today's more limited capabilities. Advancements in AI creativity and reasoning could allow GANs to generate highly infectious viral misinformation built upon complex false narratives that appeal to psychological vulnerabilities in certain demographics.[H] The algorithms could iteratively test and refine such "viral memes" to maximize their spread.

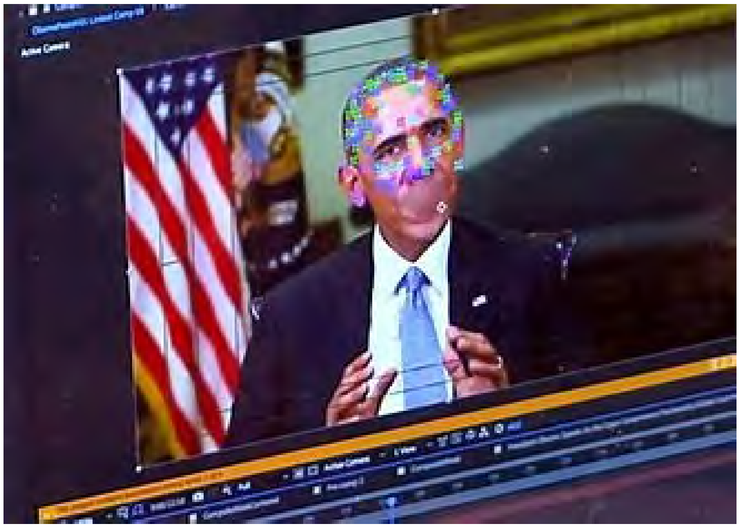


Figure 2 Researchers use AI to make realistic fake video of Obama speaking. Click on picture or go to: *https://www.businessinsider.com/ai-video-editing-fake-obama-talking-doctored-footage-2017-7* to view video. Source: BuisinessInsider.com

According to Gartner's AI development predictions, by 2033, generative AI will reach 100x to 1,000 × scale and efficiency at a low cost, enabling new mainstream applications across industries and increasing opportunities for misuse.[H] As this technology proliferates, malicious deployment of GANs for deception campaigns is likely to rapidly increase.

Countering these AI-enabled threats requires policy changes by technology platforms, government agencies, and militaries to detect and limit the spread of algorithmically generated fake content. It will also require the establishment of new types of digital forensic capabilities to authenticate media and safeguard organizations, as well as public discourse. Although GAN algorithms hold tremendous promise for commercial industries, their progression also heralds unpredictable threats on the horizon for our military. Tracking their evolution and actively developing countermeasures will become increasingly important in the coming years.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. The judgment relies primarily on open-source research publications regarding the capabilities of GAN systems and their rate of advancement as well as analysis from technology policy think tanks on likely implications. While these technologies forecast reports come from reputable subject matter experts and tend to corroborate them, there are inherent limitations in predicting future AI developments. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new information.

Author: CDR Robert V. Liberato

China Very Likely to Utilize Deepfakes to Attempt to Influence the 2028 United States Presidential Election

Executive Summary

It is very likely (80-95%) China will introduce deepfakes during the 2028 United States presidential election cycle to ensure the winner will be most favorable toward China advancing its national interests. China used this tactic to influence the January 2024 presidential election in Taiwan, but was unsuccessful in preventing a pro-independence candidate from winning. The United States National Intelligence Council published a report addressing China's involvement in influencing the 2022 United States midterm congressional elections. United States officials are also concerned about China's involvement in the 2024 United States presidential election, despite the Chinese Embassy denying all allegations and their efforts not having the intended effects of swaying the election in the Chinese Communist Party's favor.

Discussion

Leading up to the Taiwan presidential election in January 2024, China engaged in a disinformation campaign to confuse Taiwanese voters and make them question the legitimacy of their election.[M] One of the tools used by China was deepfake videos.[M] The Chinese Communist Party (CCP) also produced a 300 page book, and had artificial intelligence generated newscasters read parts of it on social media.[H] Figure 1 shows a clip from a video of Taiwan's Foreign Minister, Joseph Wu, addressing the CCP efforts to promote disinformation. Ethan Tu, founder of Taiwan AI Labs, indicated the misinformation attacks have increased since 2018 and will continue to do so.[M] Tu gave credit to the effectiveness of this tactic because people fear speaking out against it.[M] However, China's misinformation campaign did not have the desired effect of electing a more preferable candidate to Beijing, one who does not support Taiwan independence.[H]

Figure 1 Joseph Wu, Foriegn Minister of Taiwan addressing Chinese attempts to influence Taiwan's 2024 election. Video at: https://www.wsj.com/video/taiwan-fights-onslaught-of-chinese-disinformation-ahead-of-key-election/F6F0C406-164B-4572-86B0-F54736B2B163 Source: Wall Street Journal.

The United States Office of the Director of National Intelligence indicated China's senior leaders increased efforts to influence United States policy in the 2022 United States midterm elections.[H] A declassified intelligence report indicated China's intention to target the elections of specific members of Congress due to their anti-China views.[M] China's purpose was to increase the probability of pro-China candidates being elected.[H] The report also asserted the People's Republic of China, Chinese diplomats, and online actors were all involved in these influence activities.[H] Liu Pengyu, spokesperson for the Chinese Embassy denied the allegations from the report.[H]

United States Senator Pete Ricketts voiced his concerns about China's artificial intelligence operations playing an influence role in the 2024 presidential election.[H] Meta, the company which owns Facebook and Instagram, banned thousands of Chinese Facebook accounts for impersonating American accounts and posting fake news to influence the 2024 elections.[H] Meta's Global Threat Intelligence Lead, Ben Nimmo, noted an increase in Chinese online influence operations since 2020.[H] Doublethink Lab is a company which "aims to track and counter Chinese disinformation aimed at Taiwan.[H] Doublethink Lab's acting director, Lennon Chang, indicated the United States could experience the same type of influence operations in the 2024 presidential election that Taiwan faced in January 2024.[H] Given the recent history of the CCP attempts to influence elections, the increase of Chinese social media accounts, and growing concerns in congress, it is very likely the CCP will continue their misinformation attempts in the 2028 United States' presidential election.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and corroborated the information despite news sources having different political leanings. Much of this information is new and ongoing. Adequate time was allowed to prepare this report but the analyst worked alone and did not use a structured method. This report is unlikely to change prior to the end of the forecast due to a lack of international norms and regulations concerning online misinformation which will attempt to sway voters.

Author: LTC Charles Moss

The Proliferation of Disinformation Bots By Adversaries Against the United States Military is Very Likely By 2033 Due to Improved AI Capabilities

Executive Summary

The use of disinformation bots by adversaries against the US military is very likely (80-95%) by 2033 due to rapid improvements in artificial intelligence (AI) capabilities for generating synthetic text, audio, video and imagery. Despite countermeasures, the increasing sophistication of AI will enable adversaries to create highly convincing fake content to deceive and disrupt US operations. Recent demonstrations of deepfake technology indicate adversaries are rapidly acquiring the capability to generate highly realistic fake content. Paired with increasing sophistication of bot networks, this points to an extremely high likelihood of weaponization by 2033.

Discussion

Generative AI models can now create realistic fake text, images, and videos. This has enabled adversaries to weaponize these models for disinformation campaigns that aim to interfere with elections, erode public trust, and deepen social divisions.

A disinformation bot is a type of automated software program that spreads false or misleading information online without human intervention.[H] Some key characteristics of disinformation bots include[M]:

- Automatically generates and spreads intentionally false content to manipulate public perception
- Poses as real users on social platforms, making them hard to distinguish
- Operates at high volumes and speeds to flood platforms before moderators can respond
- May be coordinated in networks to create illusion of grassroots support for false narratives

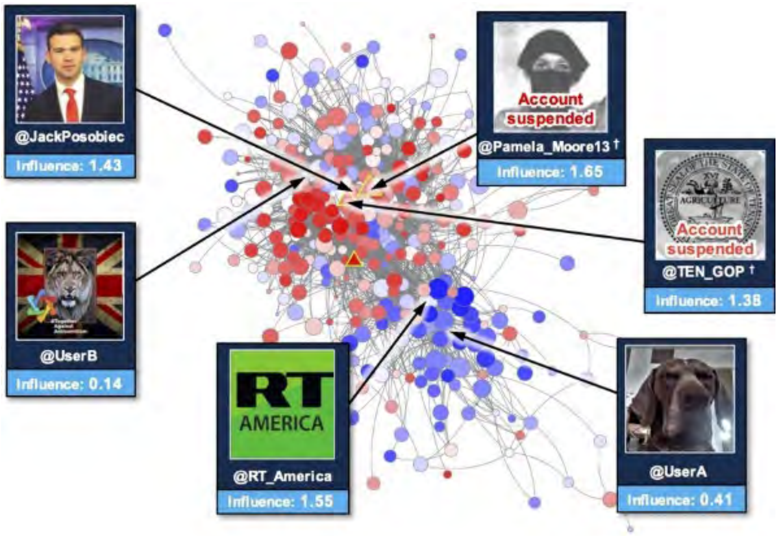


Figure 1 The Reconnaissance of Influence Operations software generates a "tree" that depicts the scope of a retweet network. Source: https://www.ll.mit.edu/r-d/projects/reconnaissance-influence-operations

Recent examples demonstrate how adversaries have already deployed disinformation bots to target military operations. During the 2022 Russian invasion of Ukraine, bots promoted false narratives around Russian troop positions to confuse Ukrainian forces regarding the direction of assaults. Russian bots also spread doctored videos and directives pretending to originate from the Ukrainian military leadership in efforts to goad forces into ambushes. Their automation and coordination make them powerful tools for manipulating public perceptions. Identifying and mitigating their harmful effects continues to be a major challenge facing online platforms, governments and the general public.

Key adversaries, including Russia, China, Iran, and North Korea, have highly sophisticated cyber programs and have already demonstrated an interest in using AI-enabled disinformation tactics.[H] Many have exploited the internet to spread propaganda and disinformation to weaken their competitors. For example, Russia's official military doctrine calls to "exert simultaneous pressure on the enemy throughout the enemy's territory in the global information space".[H] These adversaries are very likely to have enormous potential for leveraging AI-powered bots to spread targeted disinformation within US military channels. In addition to nation-state adversaries, terrorist groups and other non-state actors have strong incentives to utilize disinformation bots to sow confusion without attributable links back to them.

The accelerated pace of AI research in generative models points to a future capability for adversaries to flood US military personnel with hyper realistic fake content that spreads rapidly through social platforms and messaging apps. This could include fake troop movements, false combat scenarios, and misleading directives aimed at degrading coordination and trust in the operations.

Of principal concern is Russia, which combines advanced AI and cyber capabilities with demonstrated willingness to utilize disinformation against military targets.[H] China also runs vast online propaganda operations and presents increasing Risk from its rapidly developing AI programs.[M] Iran and North Korea present additional, but likely more limited near-term threats of employing sophisticated disinformation bots against US military channels.

Over the past decade, several teams have sought to develop algorithms that successfully identify bots online. MIT Lincoln Laboratory is developing a Reconnaissance of Influence Operations (RIO) AI system (Figure 1) to understand and automatically detect disinformation narratives, as well as those individuals who are spreading the narratives within social media networks.[H] RIO will also have the ability to help those using the system to forecast how different countermeasures might halt the spread of a particular disinformation campaign. A follow-up program is also underway to dive into the cognitive aspects of influence operations and how disinformation affects individual attitudes and behaviors.[M]

Additional defensive measures will be necessary to counter increasingly advanced disinformation bots in the future. These could include cyber operations to take down adversarial bot networks and improved identity verification procedures to prevent automated accounts from penetrating US military channels under the guise of authentic personnel. Ongoing training for military members on responsible information sharing and heightened vigilance against suspicious directives will also prove critical in this emerging threat landscape. While some detection systems are in early phases, adversarial machine learning techniques could enable disinformation bots to evade these defenses.

Analytical Confidence

Confidence level is *moderate* in the core judgment that disinformation bots will continue rapidly proliferating over the next 2-4 years. This assessment is based on clear evidence regarding the direction of technology developments enabling more advanced bots as well as unambiguous profit motives driving their creation despite ethical issues. Regulatory inability to constrain growth in the near term solidifies this outlook. However, some uncertainty exists around potential future actions by social media platforms if issues escalate further.

Author: CDR Robert V. Liberato

United States Open-Source Intelligence Reporting Will Likely Contain Artificial Intelligence Generated Information by 2033

Executive Summary

The proliferation of Artificial Intelligence Generated Content (AIGC) makes it likely (55-80%) that by 2033, open-source intelligence (OSINT) collectors will unwittingly collect and disseminate AI-generated information. Sophisticated adversarial AI (AAI) networks will likely evade attribution, resulting in AIGC entering intelligence reporting. Without AI policies and frameworks, it is very likely (80-95%) that the intelligence community's (IC) OSINT collection cycle will be unable to attribute AI-generated information before it enters intelligence reporting.

Discussion

AI-generated information will likely enter IC reporting through OSINT reporting. According to researchers at the Center for Digital Ethics and the *Social Science Research Network* (SSRN), OSINT comprises upwards of 80 percent of reporting for intelligence agencies and law enforcement.[H] AIGC news and information proliferates online and will almost certainly (95-99%) be collected for OSINT reporting. In May 2023, media watchdog *News Guard* identified 49 news and media sites that were solely AI-generated.[H] Other professional fields, such as legal services, also grapple with the complexity of attributing AIGC as a source or its implications for the admissibility of evidence.[H] The proliferation of AIGC online increases the probability of unwitting collection and dissemination of AIGC into the IC.

CSIS
Open-Source, Cloud-Based, AI-Enabled Repo...
Watch later
Share
Move Over JARVIS, Meet OSCAR
Open-Source, Cloud-Based, AI-Enabled Reporting for the Intelligence Community
MORE VIDEOS
CSIS
0:01 / 2:13
YouTube

Figure 1: Tutorial on OSCAR capabilities. Click on photo or go to https://youtu.be/hv5UXPe4zUo Source: CSIS

Attributing collected information to adversarial AAI networks is a growing national security concern. The Department of Homeland Security (DHS), in its 2023 "Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats" report, described an AAI network as a significant AI threat to the United States.[H] The DHS noted concern about a state-sponsored AAI's ability to create "…realistic human-like text, [which] opens the door for malicious actors to be able to engineer open-source information campaigns at scale." [H] Peer competitors, like China, consider the information environment a battlefield and are incorporating AIGC into information warfare.[H, H] China relied on large language models (LLM) AAIs to influence the 2024 Taiwan elections by producing high-quality news videos and AI-generated bots to sway Taiwanese opinion.[M] Russia also uses AAIs to create open-source media content to impact national security decisions. Russian attributed AI media articles influenced some members of Congress regarding aid for Ukraine through convincing news stories that Ukrainian President Volodymyr Zelensky had used United States Ukraine aid to purchase a personal yacht.[H] It is very likely that by 2028, adversaries will continue to use AAI networks to target the United States OSINT collection system with open-source media sites.

The volume of data collection and reliance on AI alone to sort the data will almost certainly contribute to AIGC entering the intelligence stream. In 2022, the Center for Strategic and International Studies envisioned an Open Source Cloud based AI-enabled Reporting, OSCAR, as a future component of OSINT Collection.[H] "OSCAR" would rely on machine learning to sort through massive amounts of data and automate collection (See Figure 1). But, as recently as March 2024, United States government officials and Internet researchers at Clemson University identified Russian-attributed media sites with genuine sounding names such as "Chicago Chronicle," "Miami Chronicle," and "D.C. Weekly,"[H] which can fool OSINT collection systems like "OSCAR." In January 2024, during a Google panel, Colonel (COL) Richard Leach, G2 for the Defense Information Systems Agency, remarked on the massive volume of information the Department of Defense (DoD) processes for decision-makers and that individual analysts do not have time or capacity to analyze all of the data and must rely on AI for data management.[M] During an interview on 9 February 2024, Dr. Herb Lin, a senior research scholar at Stanford University, when queried about AIGC entering intelligence reporting, stated that AI familiarization, not technology, will very likely prevent AIGC from impacting intelligence reporting.[10] Without DoD frameworks that integrate AI and OSINT and require training for humans in the OSINT collection cycle, fully autonomous AI collection will very likely enable AAI to enter IC reporting.

The decentralized nature of the United States' OSINT methodologies and lack of centralized policy make it almost certain that the DoD cannot withstand an AAI threat to OSINT collection. Microsoft's February 2024 report on AAI using LLMs to spread misinformation and create content affirms the enduring threat of AAI towards OSINT collection.[H] The private sector OSINT company FIVECAST details specific rules and principles that govern

[10] Herb Lin, Interview with Dr. Herb Lin Regarding Artificial Intelligence and the Psycho-Cognitive Domain, Teams Interview, February 9, 2024.

its OSINT collection and attribution to prevent AIGC and AAI attacks.[M] In contrast, the Government Accountability Office's (GAO) 2021 *Defense Intelligence and Security* report to Congress determined that the DoD's OSINT policy documents "…governing the OSINT mission area do not currently include short- or long-term outcomes, nor metrics tied to outcomes tracking performance and accountability" and that the "…DIA's OSINT authorities and responsibilities to set OSINT standards and requirements are unclear and not well understood."[H] The GAO's 2022 report on AI and the DoD outlined several areas for improvement, such as policies on AI and the need for "…guidance that clearly defines the roles and responsibilities of components that participate in AI activities."[H] A 2024-2026 OSINT strategy guidance by the Director of National Intelligence concluded that AI will play a critical role in OSINT collection, and OSINT collection training must be updated to account for AIGC[H], yet the DoD's OSINT collection frameworks are not stable enough to permit full automation of OSINT collection. Without updated training, frameworks, and tradecraft, it is almost certain the DoD OSINT enterprise is not prepared with the experience, policies, and strategies to detect and attribute AAI and prevent AIGC from polluting intelligence reporting.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were reliable and tended to corroborate one another. The author personally interviewed Dr. Herb Lin for the project. Perplexity and ideas from the results were used in further research. Perplexity suggested sources that were validated and then used as references. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change due to new developments, such as the rapid implementation of OSINT policies, a joint OSINT office within the DoD, and/or the establishment of a specific OSINT career field for civilians and military. This report is also sensitive to the future of OSINT collection with the DoD and, due to the rapidly evolving nature of AIGC, if the DoD would shift OSINT collection to government contractors.

Author: Mr. Tom M. Jackson

Generative AI Highly Likely to Drive Evolution in Targeted Cyber Espionage Targeting Methodology from 2025 to 2030

Executive Summary

The introduction of AI and AI Generated Content into existing cyber capabilities is highly likely (80-95%) to pose an increased risk to corporate and industry leaders between 2025 and 2030 despite workplace cyber protections. Cyber spies will likely use AI Generated Content to bypass existing cyber security creating a window where corporate and industry leaders will be vulnerable to exploitation until security practices catch up.

Discussion

The introduction of Artificial Intelligence Generated Content (AIGC) into cybercriminal and state sponsored cyber espionage is highly likely to drive a change in the populations most vulnerable to cyber-attacks. Currently, the most vulnerable phishing targets are young adults and adults over 75.[H] These populations tend not to benefit from institutional protections afforded by workplace cyber security practices. However, leveraging AI and AIGC in a two-step process will bypass most existing institutional security through a targeting approach.

The combination of an increasingly digitized workforce and AI enabled threat results in an opportunity for threat cyber espionage operations. Prior to 2020 companies were already increasingly digitizing and Covid drove more companies to operate in a digitally enabled cloud-based architecture.[H] By late 2025 Cyber criminals and state sponsored cyber threats will be positioned to exploit a physically isolated workforce that is increasingly reliant on cloud infrastructure and working with collaborative technologies in novel ways.[M]

Combining AI into spear phishing cyber espionage operations will undoubtedly enable a shift to corporate middle management as a new vulnerable population despite existing cyber protections.[M] Phishing is currently the most prolific and revenue generating, rising trend in cyber-criminal activity.[M] In the third quarter of 2022, alone, The Anti-Phishing Working Group (APWG)

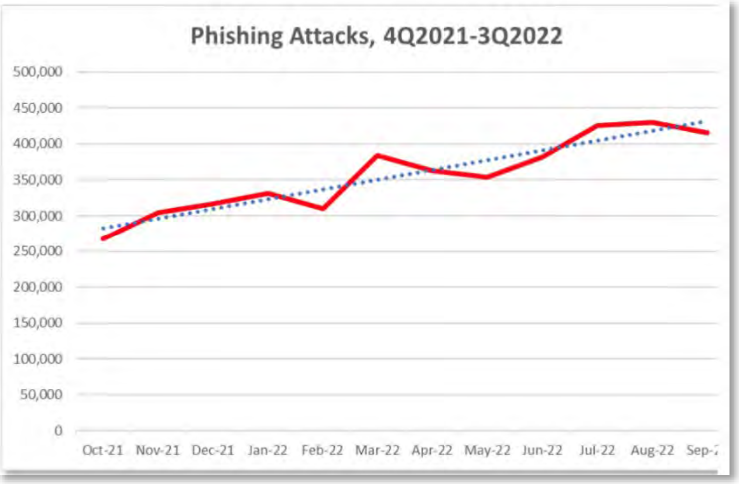


Figure 1 Phishing Attacks on the Rise. Click on picture or go to: https://apwg.org/. Source: Anti-Phishing Working Group.

comprised of industry leading cyber protection professionals observed 1,270,883 total phishing attacks (see figure), a new record and the worst quarter for phishing that APWG has ever observed.[H] If current trends continue, it is inevitable that states adopt phishing as part of state cyber espionage to gain access to networks or intellectual property.[H] AI enabled phishing and cyber espionage operations will likely enable threat actors to move away from data breaches into more sophisticated operations.

The first step in developing and AI enabled cyber espionage capability is collection key data on target populations. AI is capable of polling vast data and media repositories to collect and synthesize publicly available data, in fact this is how many AI's are trained.[M] Social Canvassing AI's can then focus on target populations such as those of corporate or industrial leaders, searching for relevant data. The second step is to use Generative AI to create content tailored to specific individuals. The key is to generate realistic content that appeals to the target. The goal of this approach is to gain access to networks and intellectual property.[M] The net result is a socially engineered targeted approach is that cyber operatives are invited into the network by increasingly realistic content. [M] A next generation use of AIGC is on the near frontier and uses generative AIs to exploit stolen credentials to replicate the target within the network based on available data.

The emerging nature of AIGC will likely (55%-80%) make the window of opportunity to exploit the workforce fleeting. AIGC and associated capabilities are still emerging and will likely result in novel uses in cyber espionage events. The rush to use will be balanced by states' desire for sophistication to avoid detection. Therefore, the window for maximum effect will likely be between 2025 and 2030. It is likely that industry will develop controls and protections for its workforce but those protections will take years to fully implement. Unfortunately, by the time emerging protections are in place, hostile states' cyber espionage operatives will already be in the network.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another, however most focused on current security conditions. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is subject to change based on the pace at which cyber protection capabilities designed to counter AIGC are implemented at scale.

Author: COL Robert M. Richardson

Department of Defense Very Likely To Use Intelligence Community and U.S. Gov Agencies in lieu of Private Sector for Artificial Intelligence Detection by 2033

Executive Summary

By 2033, it is very likely (80-95%) that the Department of Defense will rely on the intelligence community (IC) or cybersecurity departments within U.S. government agencies, rather than the private sector, for Artificial Intelligence Generated Content (AIGC) detection tools and procedures. The IC will continue to have requirements for collection and the necessity for safeguarding or controlling misinformation and disinformation. Despite the commercial sector's speed of AIGC advancements, there is limited financial incentive for individual businesses to prioritize AIGC detection software or work with the federal government on joint partnerships.

Discussion

Since the emergence of artificial intelligence, some commercial companies have been leading innovation, and successfully providing intelligence services to the government.[H] However, even companies which offer the best intelligence services are still beholden to private investors, strategic partners' agendas, or the drive for profitability.[M] While many companies do work with the IC and provide services for a cost, they are ultimately data brokers vice data protectors. Figure 1 illustrates a rendering of the Republic of Indonesia's earth Imaging constellation, resulting from a commercial (U.S.) company, BlackSky that just signed a $50 million deal to supply imagery services and satellites to Indonesia.

Figure 1: Rendering of the Republic of Indonesia's Earth imaging constellation. Credit: Thales Alenia, full story at website: https://spacenews.com/blacksky-inks-50-million-deal-to-supply-imagery-services-and-satellites-to-indonesia/

With the rise of artificial intelligence generated content (AIGC), such as deepfake pictures or videos, the information environment has changed within the open-source intelligence discipline.[M] However, checking the veracity of data given the deepfake threats is getting engrained in IC agencies.[M] Moreover, government agencies have been aware of this threat since 2021[M] so it is unlikely that detection tools or safeguarding measures have not been already put in place.

Because of the intelligence community (IC)'s mission, the requirements for IC collection cannot be compromised. Policymakers and strategic leaders demand intelligence estimates and continual indications and warnings, which are based on having raw information collected for fusion and analysis. These demands will not cease, nor decrease in the future. U.S. strategy and policy documents cement the IC's role as critical, with no-fail mission requirements, which the commercial sector does not have to consider.[H] While the commercial sector will likely demonstrate how their social media platforms can be trusted,[M] the IC will be focused narrowly on the data they need to collect.[M] Finally, the computer systems that the IC often use for collection and processing, due to specific intelligence disciplines is often on a higher classification network than what commercial companies use.[H]

The commercial sector is very likely (80-95%) to invest in the AI tools that are yielding the highest returns,[M] which may not necessarily include AIGC detection software. An expert in AIGC, Mr. Edmon Begoli, who leads the Oak Ridge National Laboratory's Center for AI Security Research has stated that commercial and private sector companies will be focused on creating the most efficient large language models for their customers, not looking at how to protect or check the veracity of data. The National Labs within the Department of Energy are not only funded and resourced for these endeavors, but work in collaboration and synchronize with many of the cyber departments within the IC.[H]

Analytic Confidence

The analytic confidence for this estimate is *moderate*. This estimate is focused on a potential change in policy, should the U.S. government consider using more commercial or private sector AIGC during the intelligence collection. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change. Intelligence reporting at a higher classification level could affect this forecast.

Author: LTC Katherine M. Ogletree

By 2029 U.S. Government Attribution Tools Very Likely to Use Machine Learning

Executive Summary

By 2029, it is very likely (80-95%) that U.S. government attribution methods in cyber environments will utilize machine learning (ML), a type of artificial intelligence. This technique would be a result of consolidated collaboration between the commercial sector, academia, and government after successfully testing such tools in government laboratories with proven success rates. Using these automated forensic tools seeks to increase efficiency and speed while being able to maintain a very detailed focus for specific adversarial modus operandi. Comparing how adversaries conduct cyber operations will likely assist with attribution or detection, especially across other AI generated content, like deepfakes. Despite AI hallucinations and biases, the persistent feedback loop will enhance ML production and showcase its value as a means for identifying attribution.

Discussion

Attribution models seek to determine the responsibility for an action, which has become increasingly more difficult in the information sphere. The U.S. government has a vested interest in this topic, considering both China and Russia have been using cyber to conduct information operations, such as disinformation, as well as conduct espionage, intelligence gathering, and targeting U.S. states' critical infrastructure.[M]

Figure 1: An example of using deep machine learning to detect deepfakes. This company, IEEE Xpert, is a relatively new engineering private firm in India, dedicated to Research and Development, learn more here https://www.ieeexpert.com/

Source attribution is critical to hold malicious actors accountable for their actions, especially if cyber actions are conducted in conjunction with elements of warfare, as Russia has been doing in its conflict with Ukraine.[H] Attribution is just as important when not in direct conflict, especially when the implications result in critical infrastructure getting shut down as in the Colonial pipeline attack.[H] Given the rise of cybercrime, it is understandable that the commercial and private sectors are just as concerned as the U.S. government with attribution.[M]

These aligning interests are likely (55-80%) the main reason the Intelligence Advanced Research Projects Activity (IARPA) is harnessing the power of machine learning, a type of artificial intelligence[M], to pinpoint cyber attribution. In October 2023, the Office of the Director of National Intelligence pitched a proposal's day briefing to industry, both commercial and private sectors, titled "Securing our Underlying Resource in Cyber Environments (SoURCE).[M] The program is seeking novel technologies that will make determinations of the most likely attackers, based on coding styles, and measure the similarity between files to provide likely origins (country, groups, individuals, etc.).[M] Testing and evaluation partners include Sandia National Laboratory, Lawrence Livermore National Laboratory, and the Software Engineering Institute.[M]

The Department of Homeland Security is already testing out detection algorithms to scrutinize digital context for inconsistencies and manipulation.[H] Figure 1 is an example of how an Indian company, specializing in research and development for engineering, explains the concept of using deep machine learning to unmask deepfakes.[M]

Despite known AI hallucinations and bias,[M], the ability to create a feedback loop will enhance AI/ML into the attribution process. In machine learning, feedback loops provide an opportunity to correct a model's decision and subsequently enhance the output going forward.[M] [M] Continuous improvement with AI/ML will make U.S. government agencies more likely to adopt tools in their attribution processes.

Analytic Confidence

The analytic confidence for this estimate is *moderate*. Sources were generally reliable and tended to corroborate one another. No AI tools, other than Grammarly, were used in this estimate. There was adequate time, but the analyst worked alone and did not use a structured method. Furthermore, given the lengthy time frame of the estimate, this report is sensitive to change. Intelligence reporting at a higher classification level could affect this forecast.

Author: Katherine M. Ogletree

ERGO SUM

MACHINA

Annexes

Annex A – Terms of Reference

Terms of Reference: Weaponization of Artificial Intelligence: Effects and Responses

For:

LTG Laura Potter
Deputy Chief of Staff, G2

By:

Team Ergo Sum Machina USAWC

November 21st, 2023

Terms of Reference:
Weaponization of Artificial Intelligence: Effects and Responses

Requirement:

How might future adversaries use emerging AIGC tools, techniques, and processes for deception and manipulation of the psycho-cognitive domain through 2033?

- Sub 1) What are the likely effects and implications for the use of emerging AIGC tools, techniques, and processes?

- Sub 2) What new detection, attribution, and countermeasure processes will need to be developed to counter the threat of AIGC?

Methodology:

- **Gathering Phase-December 2023 to January 2024:** The team intends to focus on a qualitative over quantitative approach. The team has access to various sources, such as academic and open-source research, interviews, podcasts, and Artificial Intelligence (AI) enabled open-source data gathering. During this phase, the team will consider areas of research emphasis, such as:

  - Research and understand the primary AI Platforms.
  - Explore historical analysis on the impact of AI in psycho-cognitive space.
  - Explore existing AI detection mechanisms.
  - Research and review current private sector detection and countermeasure best practices.
  - Explore research on AI risks in the technical space.
  - Discover if the Department of Defense (DoD) is working on AI-related projects.

- **Analyzing Phase-January 2024-February 2024:** The team will consider a cross-functional approach and an investigative/analytical phase line that would:
  - Initially consolidate the information and assess value based on grouping, filtering, prioritizing, etc.
  - Evaluate the impact of AI in the psycho-cognitive and technical space.
  - Evaluate how the findings inform the near future (1-5 years) or the future (10 years).

  - Examine the impact of the findings for each domain to identify the most at-risk domain and future vulnerabilities.
  - Explore future AI inflection points juxtaposed to likely countermeasures.

- **Compiling Phase-March 2024-April 2024:** Compile concepts and prepare report (March 2024)
  - Prepare a comprehensive report that includes the teams' findings regarding adversaries' attempts to shape the psycho-cognitive domain and anything the US Army can employ to counter this.
  - Create a briefing or any visual aids (potential modeling or demonstration) accompanying the report.

- **Final Briefing:** The team will brief the G2 and their staff on/about April 2024.

Challenges:

- **Time:** This project will run concurrently with a full graduate course load and an extra elective.
- **Resources:** Limited funding is available to support travel and other related expenses.
- **Future Leaning Focus:** Because this study is extrapolatory, some assumptions will be included to forward the research.
- **Limited Information Streams:** Equipment and resources constrain the team to an open-source environment so that the final product will be Unclassified.
- **Speed of Transformation:** AI research, platforms, and capabilities are accelerating exponentially. The data gathered this year may be irrelevant or stale next year.
- **Complexity:** AI is a new area of study for the layperson, and the team lacks the expertise or domain knowledge beyond the basic user level in the field. While the team is comprised of educated and experienced individuals, AI, in general, presents a technical challenge.
- **Language:** Research on adversaries may require translation as no team member is fluent in Chinese or Russian.
- **Regulatory:** There is limited (or no) international regulatory guidance when using artificial intelligence and machine learning tools, thus adversaries may already be incorporating AI effects into their wartime planning.

Resources:

- **Personnel:** The team is comprised of a combination of Army and Navy, bringing a wealth of experience across domains.
- **Institutional:** The team has access to a significant library of journal resources, AI websites, a writing laboratory, cognitive spaces, and an adaptive learning environment at the Army War College (AWC).
- **Funds:** The AWC Futures program has a temporary duty and production budget.
- **Technology:** The AWC has access to a Futures Lab, an organic chatbot, and in-house scientific assistance from technologists.
- **AI Tools:** The team has access to numerous AI tools at low or no cost to support experimentation and inform observations from the field.
- **Networks:** The team can leverage extensive DoD and private sector networks to identify subject-matter experts for interviews and assistance.

Administration:

- The final product will be delivered in PDF format and is for the sole use of LTG Potter, DA G2, her team, and any personnel she chooses to designate.
- Contact information:
    - Team Point of Contact:
        - COL Robert “Marc” Richardson, robert.m.richardson38.mil@armywarcollege.edu, 253-653-4969.
    - Alternate Point of Contact:
        - Mr. Tom Jackson, tom.m.jackson.civ@armywarcollege.edu, 862-276-7986.
    - Team Members:
        - CDR Robert “Rob” Liberato, robert.v.liberato.mil@armywarcollege.edu, 808-339-0605
        - LTC Charles Moss, charles.c.moss.mil@armywarcollege.edu, 901-289-0181.
        - LTC Katherine “Kate” Ogletree, katherine.m.ogletree.mil@armywarcollege.edu, 757-945-1835.
        - Team Physical Mailing address: 651 Wright Avenue, Carlisle Barracks, PA, 17013

Annex B – Assessing Analytic Confidence

The analysts were not subject matter experts and some topics required extensive scientific knowledge to fully grasp. The analysts worked independently and collaboratively to answer the question. They utilized a combination of structured analytic techniques including nominal group technique and network analysis. The team evaluated their analytic confidence using Peterson’s Analytic Confidence Factors coupled with the Friedman Corollaries.

Peterson’s Analytic Confidence Factors

- How reliable are the sources?
- How well do the independent sources corroborate each other?
- What is my/my team’s level of expertise?
- How effective was my analytic collaboration?
- Did I use any structured techniques in my analysis?
- How difficult did I perceive the task to be?
- Did I have enough time to complete the task?

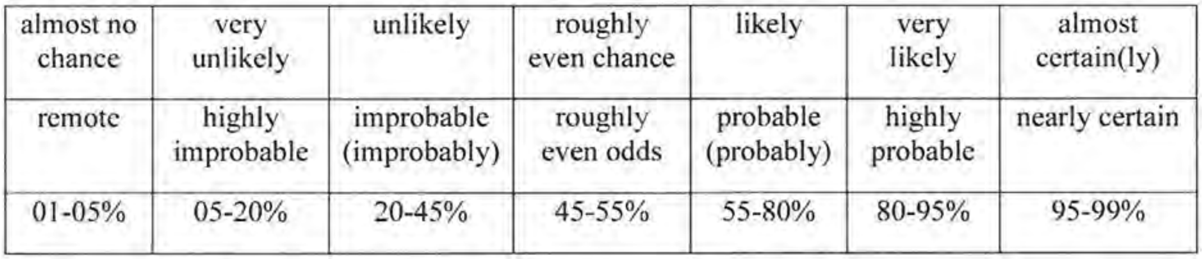
Friedman Corollaries

- Is my estimate within the range of reasonable opinion surrounding the question?
- How likely is it that new information will change my estimate?

Annex C – Words of Estimated Probability

Team Ergo Sum Machina utilized the Intelligence Community Directive (ICD) 203 as their guide for determining their Words of Estimative Probability (WEP) for expressions of likelihood or probability, an analytic product must use one of the following sets of terms for determining the likely applications across the continuum of AIGC through 2033.

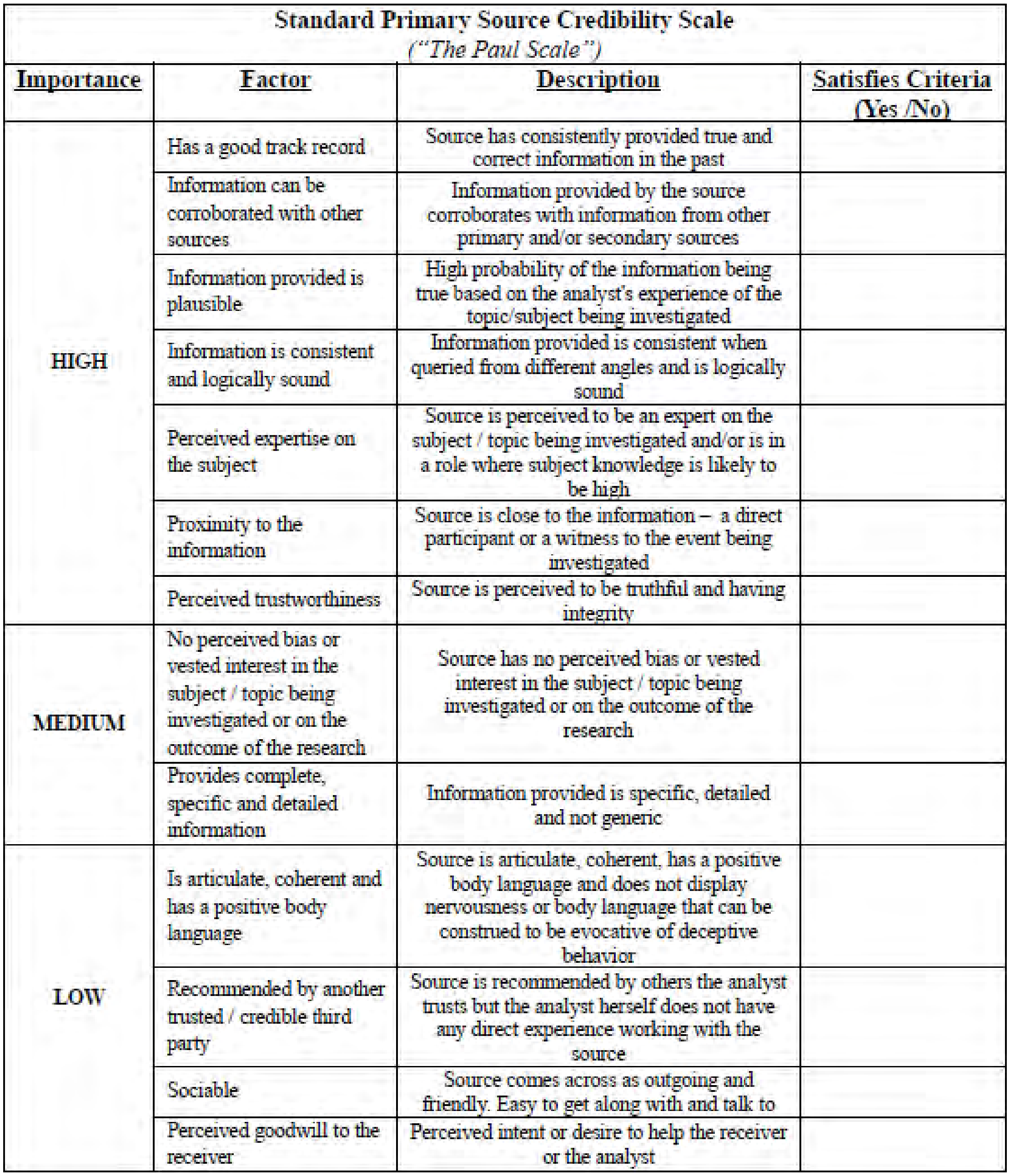
Intelligence Community Directive (ICD) 203 of Estimative Words



almost no chance	very unlikely	unlikely	roughly even chance	likely	very likely	almost certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01-05%	05-20%	20-45%	45-55%	55-80%	80-95%	95-99%

Annex D – Standard Primary Source Credibility Scale

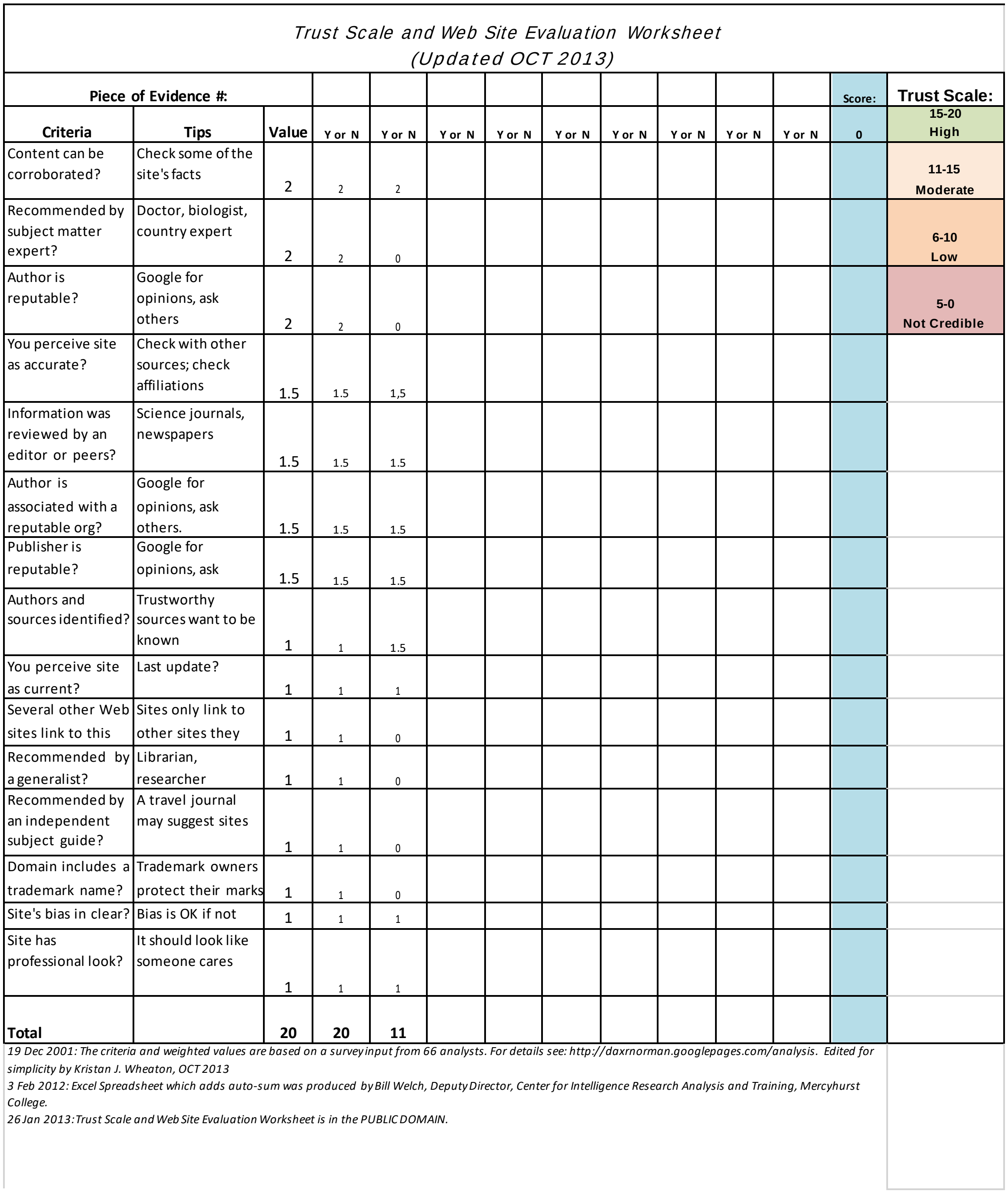
Source reliability is noted at the end of each citation as low L, moderate M, or high H. The citation is hyperlinked to the source, unless the source is a paid subscription; in that instance a footnote is provided at the end of each writing illustrating the source for credibility. Source reliability is determined using the Trust Scale and Website Evaluation Worksheet found in Annex E.



Standard Primary Source Credibility Scale
("The Paul Scale")

Importance	Factor	Description	Satisfies Criteria (Yes /No)
HIGH	Has a good track record	Source has consistently provided true and correct information in the past	
	Information can be corroborated with other sources	Information provided by the source corroborates with information from other primary and/or secondary sources	
	Information provided is plausible	High probability of the information being true based on the analyst's experience of the topic/subject being investigated	
	Information is consistent and logically sound	Information provided is consistent when queried from different angles and is logically sound	
	Perceived expertise on the subject	Source is perceived to be an expert on the subject / topic being investigated and/or is in a role where subject knowledge is likely to be high	
	Proximity to the information	Source is close to the information – a direct participant or a witness to the event being investigated	
	Perceived trustworthiness	Source is perceived to be truthful and having integrity	
MEDIUM	No perceived bias or vested interest in the subject / topic being investigated or on the outcome of the research	Source has no perceived bias or vested interest in the subject / topic being investigated or on the outcome of the research	
	Provides complete, specific and detailed information	Information provided is specific, detailed and not generic	
LOW	Is articulate, coherent and has a positive body language	Source is articulate, coherent, has a positive body language and does not display nervousness or body language that can be construed to be evocative of deceptive behavior	
	Recommended by another trusted / credible third party	Source is recommended by others the analyst trusts but the analyst herself does not have any direct experience working with the source	
	Sociable	Source comes across as outgoing and friendly. Easy to get along with and talk to	
	Perceived goodwill to the receiver	Perceived intent or desire to help the receiver or the analyst	

Annex E – Trust Scale and Web Site Evaluation Worksheet



Trust Scale and Web Site Evaluation Worksheet
(Updated OCT 2013)

Piece of Evidence #:													Score:	Trust Scale:
Criteria	**Tips**	**Value**	Y or N	Y or N	Y or N	Y or N	Y or N	Y or N	Y or N	Y or N	Y or N	0	15-20 High
Content can be corroborated?	Check some of the site's facts	2	2	2									11-15 Moderate
Recommended by subject matter expert?	Doctor, biologist, country expert	2	2	0									6-10 Low
Author is reputable?	Google for opinions, ask others	2	2	0									5-0 Not Credible
You perceive site as accurate?	Check with other sources; check affiliations	1.5	1.5	1,5									
Information was reviewed by an editor or peers?	Science journals, newspapers	1.5	1.5	1.5									
Author is associated with a reputable org?	Google for opinions, ask others.	1.5	1.5	1.5									
Publisher is reputable?	Google for opinions, ask	1.5	1.5	1.5									
Authors and sources identified?	Trustworthy sources want to be known	1	1	1.5									
You perceive site as current?	Last update?	1	1	1									
Several other Web sites link to this	Sites only link to other sites they	1	1	0									
Recommended by a generalist?	Librarian, researcher	1	1	0									
Recommended by an independent subject guide?	A travel journal may suggest sites	1	1	0									
Domain includes a trademark name?	Trademark owners protect their marks	1	1	0									
Site's bias in clear?	Bias is OK if not	1	1	1									
Site has professional look?	It should look like someone cares	1	1	1									
Total		**20**	**20**	**11**									

19 Dec 2001: The criteria and weighted values are based on a survey input from 66 analysts. For details see: http://daxrnorman.googlepages.com/analysis. Edited for simplicity by Kristan J. Wheaton, OCT 2013

3 Feb 2012: Excel Spreadsheet which adds auto-sum was produced by Bill Welch, Deputy Director, Center for Intelligence Research Analysis and Training, Mercyhurst College.

26 Jan 2013: Trust Scale and Web Site Evaluation Worksheet is in the PUBLIC DOMAIN.

Annex F – Interview/Communication Notes

Meeting Notes with Ms. Michele Flournoy
09 February 2024 @ 1300 via teams
Attendees: LTC Matt Rasmussen, Mr. Tom Jackson, COL Robert Richardson
LTC Katherine Ogletree, LTC Charles Moss, CDR Robert Liberato

Team Ergo Sum Machina met with Ms. Michele Flournoy an American defense policy advisor and former government official. She was Deputy Assistant Secretary of Defense for Strategy under President Bill Clinton and Under Secretary of Defense for Policy under President Barack Obama. Ms. Flournoy seemed like a true expert in her field. This discussion lasted approximately 40 minutes and was very productive at identifying current efforts and gaps. Below are the questions asked by the team and Ms. Flournoy's response.

How does AIGC affect the intelligence community?
Deepfakes are very real. Robocall of President Biden telling people not to vote in the primary election. Fueling an arms race towards understanding deepfakes to develop tools to identify. Adversaries will adopt a different approach. Independent social media sites will have different criteria to deny the allowance of videos.

In 2022 Russia released deepfake of Zelensky surrendering. How serious are senior leaders taking this and how are they shifting resources?
Have not seen senior leaders outside the IC combating this. Has not matured to the point of being a major resource expenditure. Technology and tool component as well as training. There is an awareness of the problem but not a strategy to prevent it.

Adversaries see deepfakes as below the zone of conflict, do you agree?
They are more aggressive offensively than the United States. We constrain ourselves more.

What concerns has the IC addressed with you?
The IC wants to identify fake content quickly to arm policy makers with a response. They want to ensure fake content isn't being used in intelligence analysis. They want to prevent the spread of fake stories.

Do you have thoughts on the IC taking the lead in this task?
Their focus will be on things they collect. The frontline will not be the IC, it will be the social media platforms who need to defend. One person's repression of deepfakes is another's repression of free speech. The IC is more focused on its own collection.

The United States is looking at laws concerning this, what can the DOD do to help?

There is value in the DOD building responsible Ai policies. The DOD needs to ensure training and education to ensure Servicemembers question sources which seem off. We have to enhance our defenses.

Following up to arms race, there is a lag between the AI tool and the detection mechanism. Will we be able to get ahead of this?
The jury is still out. The IC has cells which look at how deepfakes are produced so they can speed up detection. The bigger problem is in the commercial sector in social media platforms. I don't know what the commercial sector is doing. Where will people go when the easy uploading to social media platforms is taken away? The greatest risk is information making it to the wild, social media and media platforms. It is not a problem the government can solve alone.

Can you think of any sources or aspects of this problem we should look at?
There is a lot of work being done with spoofing. How to fool AI, work is being done my MITER, data poisoning, deception.

2-year feedback loop on effectiveness of AI, can you expound?
The head of AI for a major financial institution used AI to improve forecasting. The more feedback put into the model, improves the model. 2 years is not a magic number.

Meeting Notes with Dr. Herb Lin
09 February 2024 @ 1400 via teams
Attendees: LTC Matt Rasmussen, Mr. Tom Jackson, COL Robert Richardson
LTC Katherine Ogletree, LTC Charles Moss, CDR Robert Liberato

Team Ergo Sum Machina met with Dr. Herb Lin is senior research scholar for cyber policy and security at the Center for International Security and Cooperation and Hank J. Holland Fellow in Cyber Policy and Security at the Hoover Institution, both at Stanford University. His research interests relate broadly to policy-related dimensions of cybersecurity and cyberspace, and he is particularly interested in the use of offensive operations in cyberspace as instruments of national policy and in the security dimensions of information warfare and influence operations on national security. Dr. Herb Lin seemed like a true expert in his field. This discussion lasted approximately 40 minutes and was very productive at identifying current efforts and gaps. Below are the questions asked by the team and Dr. Herb Lin's response.

Opening remarks from Dr. Lin:
The fundamental issues will not be solved by technology. My writing was about confusion within DOD and not a technology issue. One could argument that every military operation comes down to information. Forces are creating a condition to make the other side not want to fight. It is not a question of new technology changing that. Until the DOD gets it right, new technology will not help.

Do you think changes in technology has strategic impacts?
Of course, but I don't know exactly what counts as information and what you want to do with it. The DOD goes back and forth.

Do you see the DOD as approaching information different from our adversaries?
You as the DOD have limitations on what you can do in the information realm. This is not the case in China. This is part of the Chinese plan to take advantage in the information domain. It is ridiculous that the DOD cannot look into social media as information collection. China has no constraints on being truthful.

What are your thoughts on the relationship between the DOD and industry
Social media companies are private sector entities with a profit motive?
If they do something against US interests, so what? Look at Elon Musk and what he has limited with Starlink in Ukraine. This has changed the way Ukrainians have fought. The US Reduction Act directs industries to support the US. There is no law making Starlink provide internet in Ukraine. China does have these laws. I am not advocating for the US to adopt Chinese polices. But, maybe we should have something leaning that way.

What do you think the DOD's role is in states using deepfakes to change mindset?
I wish the DOD could monitor domestic disinformation. I don't think it will for a variety of reasons because it is politically radioactive. The DOD has an interesting role in information development. There is no way of enforcing private sector to adopt them though. This is a national policy decision. I urge you to think about what the nation can do beyond the DOD purview.

Executive order 14111 advocates for action within the US, what is the impact of AIGC affecting trust of the government?
Provocative statement follows, which I don't believe, The impact of AI will not be that great. Why not? The debate is so bad and polarized that even if you multiply its affects, it cant get worse. I don't believe that, I think AI will make it worse. You know people are worried about deepfake videos but remember how many people were circulating the slowed video of Pelosi? That wasn't even a deepfake, but it had polarizing impact.

Will deepfakes change decision making?
It might but so will anything else. Liar's dividend; when there are tools available to make deepfakes, any bad information can be called a deepfake. This could negate entire an entire class of evidence. Trump is already claiming videos of him are deepfakes.

Are there any other areas I which we should look?
Look at China's 3 areas of warfare. They have principles of conflict below the level of war. They address manipulation of legal, information and a third. It is the way they think about non-kinetic warfare. The DOD doesn't understand information because it is focused on blowing things up. China's patron saint, Sun Tzu, how to win without fighting.

Are you advocating on the education side of using information in new ways?
Yes. The DOD's phasing assumes a world of peace. China's view is continuous struggle. There is no relaxation.

Meeting Notes with Mr. Edmon Begoli
29 February 2024 @ 1300 via teams
Attendees: LTC Matt Rasmussen, Mr. Tom Jackson, COL Robert Richardson
LTC Katherine Ogletree, LTC Charles Moss, CDR Robert Liberato

Team Ergo Sum Machina met with Mr. Edmon Begoli a founding director of *Oak Ridge National Lab's* (ORNL) Center for AI Security Research, and is a distinguished member of the ORNL research staff. Mr. Begoli co-leads ORNL's internal AI research initiative with focus on AI safety and security applications.He specializes in the research, design, and development of resilient, secure, and scalable machine learning and analytic architectures. Additionally, Edmon has been leading national programs focused on AI security, fraud prevention. Mr. Begoli seemed like a true expert in his field. This discussion lasted approximately 45 minutes and was very productive at identifying current efforts and gaps. Below are the questions asked by the team and Mr. Begoli's response.

Opening comments from Mr. Begoli:
Research is based on concerns of adversarial capabilities. Observations reveal a lack of advantage over our adversaries. Nation-states are racing with US to have the same computing resources. US has the most advanced chips but the gap is closing. China is major rival. Followed by Russia and middle eastern countries. ME countries are investing heavily in AI. Big nation-states will use AI in multiple ways. Chinese doctrine uses AI in all warfare domains. Second major area of concern - AI tools which are being developed can enable low-resourced groups to have access to things they wouldn't otherwise have. CBRNE threats can almost be developed by LLM. References Unknown Killer Robots on Netflix. Pharmaceutical AI can be reversed to create deadly drugs instead of healing by a simple change in coding. On misinformation - low hanging fruit, but can be deadly in some cases.

Do you have detection software and where do you see that developing?
Screen Share on AAI in Biometrics. Morphing Attacks and Evasion Attacks. Applies facial characteristics to another's face. Not detectable to humans. LLM can be told to be misinformation bots. Small changes can confuse AI and make it misclassify objects.

What did you mean by misinformation being low hanging fruit?
It is easy. It takes nothing to make it happen. AI can be trained to generate information for a variety of reasons. Referenced a deepfake of President Zalensky telling soldiers to surrender.

What are your thoughts on AI training AI?
It is a vulnerability and an opportunity. Tesla uses AI to improve automatic driving. Data poisoning is an AI problem. It makes it hallucinate. Algorithms need to be protected to

ensure they are not corrupted or manipulated. Could be manipulated to not recognize a threat and give warning.

Do you have commercial partnerships?
Companies are starting to approach us for partnership. Companies are competing for bigger models, not safer models. Companies are interested in red team efforts to find gaps in security.

Have national labs put policies in place to detect AI?
In the most important areas, yes. The federal government is paying more attention to it.

What are the emerging AI trends?
I cannot answer because Large Language Models LLM are overhyped. Many people give more credit to LLM than they are due. LLM is not mature enough to replace command and control. An important development is reinforcement learning. This teaches LLM to be more correct. Chain of Reasoning/thought is more of a command-and-control AI. 1. LLM develop further. 2. Building more effective commercial AI. 3. AI in labs are not LLM and we do not know where they are going. Human like reasoning is a goal. Another goal, AI which has physical capabilities.

What AI impacts human cognition most?
Confusion and deception today. More complicated is a dependence on AI. Not enough people know how these work. It could make us dumber by inhibiting learning.

Annex G – Acronyms/Glossary

Artificial Intelligence (AI): a broad and multifaceted field that encompasses the study, development, and application of systems, algorithms, and technologies capable of exhibiting intelligent behavior. At its core, AI aims to create machines and software that can perform tasks that typically require human intelligence

Adversarial Artificial Intelligence (AAI): the study and development of AI systems that are designed to deceive, manipulate or compromise other AI systems or machine learning models.

Artificial Intelligence-Generated Content (AIGC): The use of AI systems and technologies to create various forms of content, such as:

> Text Generation: AI language models can be used to generate human-like text, including articles, stories, poems, scripts, and even code.
>
> Image Generation: AI-powered image generation models, such as Generative Adversarial Networks (GANs) and Diffusion Models, can create realistic and novel images from textual descriptions or other input.
>
> Audio and Music Generation: AI systems can be trained to generate audio, music, and even synthetic voices by learning from existing datasets.

Audio and Music Generation: AI systems can be trained to generate audio, music, and even synthetic voices by learning from existing datasets.

Cyber Infrastructure: The integrated system of software, hardware, and human expertise that enables advanced data-intensive scientific research and discovery. SCADA systems are a component of cyberinfrastructure.

Epistemic Agency: An individual's capacity to actively and consciously engage in the process of knowledge acquisition, evaluation, and application. The ability to ascertain or attribute truth.

Image Generation: AI-powered image generation models, such as Generative Adversarial Networks (GANs) and Diffusion Models, can create realistic and novel images from textual descriptions or other input.

Large Language Model (LLM): A type of artificial intelligence (AI) system that is trained on a vast amount of text data to develop a deep understanding of language and the ability to generate human-like text.

Psycho-Cognitive Domain (PCD): The interconnected aspects of human psychology and cognition that influence an individual's thoughts, emotions, behaviors, and overall mental processes. The effects on the psycho-cognitive domain as those things that would influence the will of a service member not to do a needed action or the will of the American people not to support the war fight.

Supervisory Control and Data Acquisition (SCADA): A system that is used to monitor and control industrial processes and infrastructure in real-time. SCADA systems are widely used in industries such as manufacturing, energy, transportation, water and wastewater management, and other critical infrastructure sectors. They play a crucial role in improving operational efficiency, safety, and reliability by providing real-time visibility and control over complex industrial processes.

Synthetic Data: Artificially generated data that is designed to replicate the statistical properties and characteristics of real-world data, without directly using or replicating the original data sources. Data produced by existing AI models to generate the massive volume of high-quality, unbiased, cheap data needed for training other AI models.

Text Generation: AI language models can be used to generate human-like text, including articles, stories, poems, scripts, and even code.

Annex H – Team Ergo Sum Machina Briefing Presentation

WEAPONIZING ARTIFICIAL INTELLIGENCE GENERATED CONTENT

ERGO SUM MACHINA

USAWC Futures Seminar

COL "Marc" Richardson

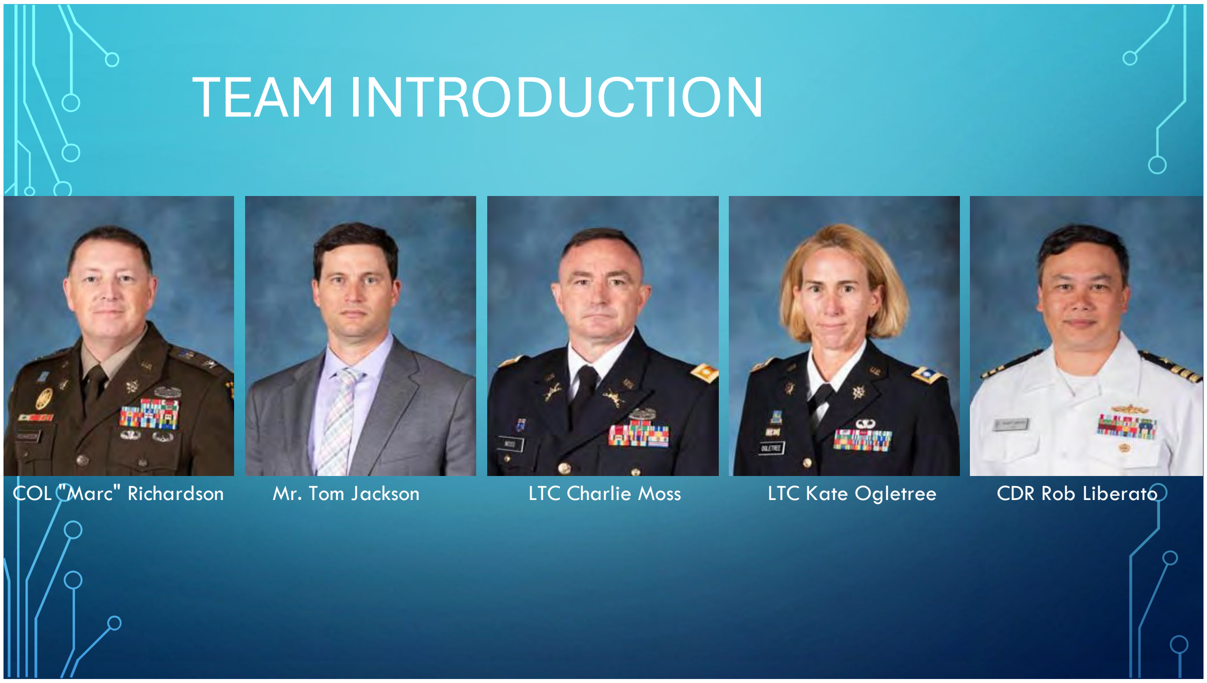
Mr. Tom Jackson

LTC Charlie Moss

LTC Kate Ogletree

CDR Rob Liberato

ICOD: March 2024



TEAM INTRODUCTION

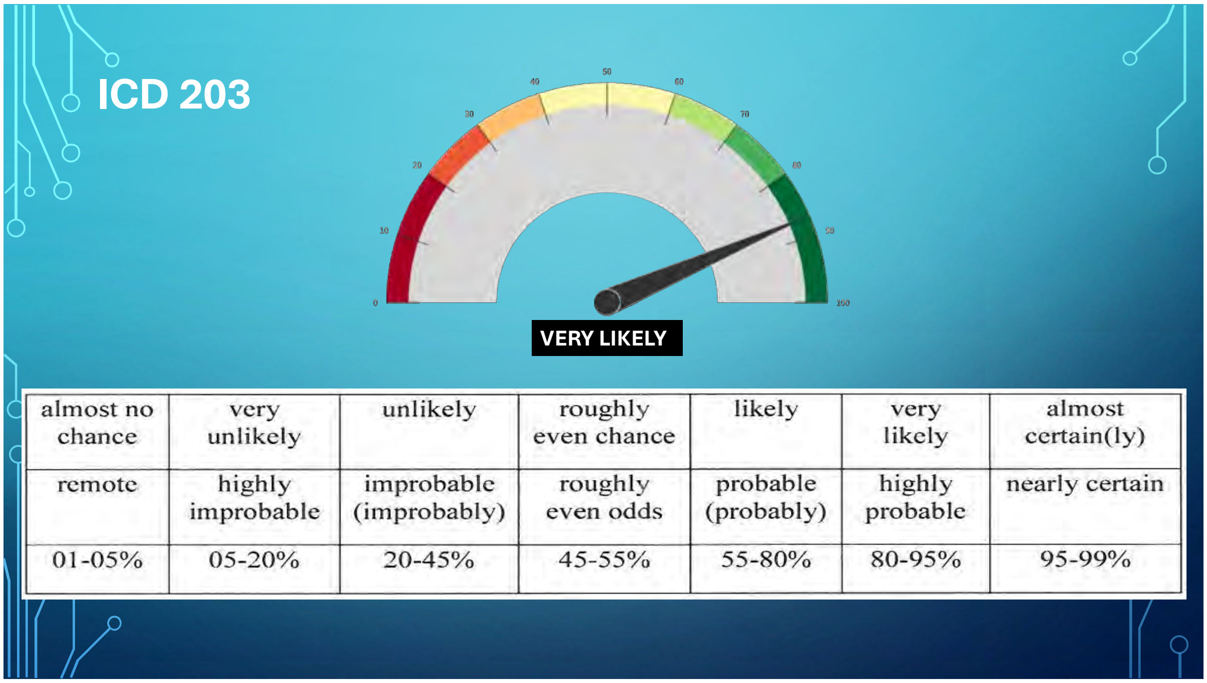
COL "Marc" Richardson

Mr. Tom Jackson

LTC Charlie Moss

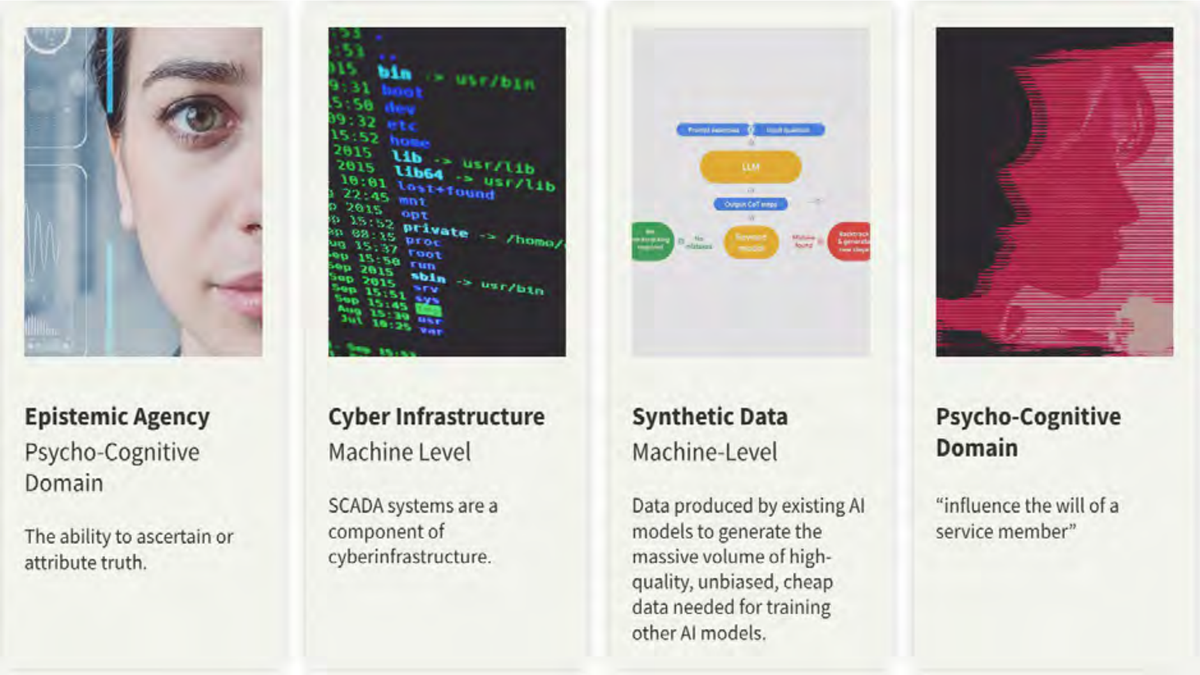
LTC Kate Ogletree

CDR Rob Liberato



ICD 203

almost no chance	very unlikely	unlikely	roughly even chance	likely	very likely	almost certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01-05%	05-20%	20-45%	45-55%	55-80%	80-95%	95-99%



Epistemic Agency
Psycho-Cognitive Domain

The ability to ascertain or attribute truth.

Cyber Infrastructure
Machine Level

SCADA systems are a component of cyberinfrastructure.

Synthetic Data
Machine-Level

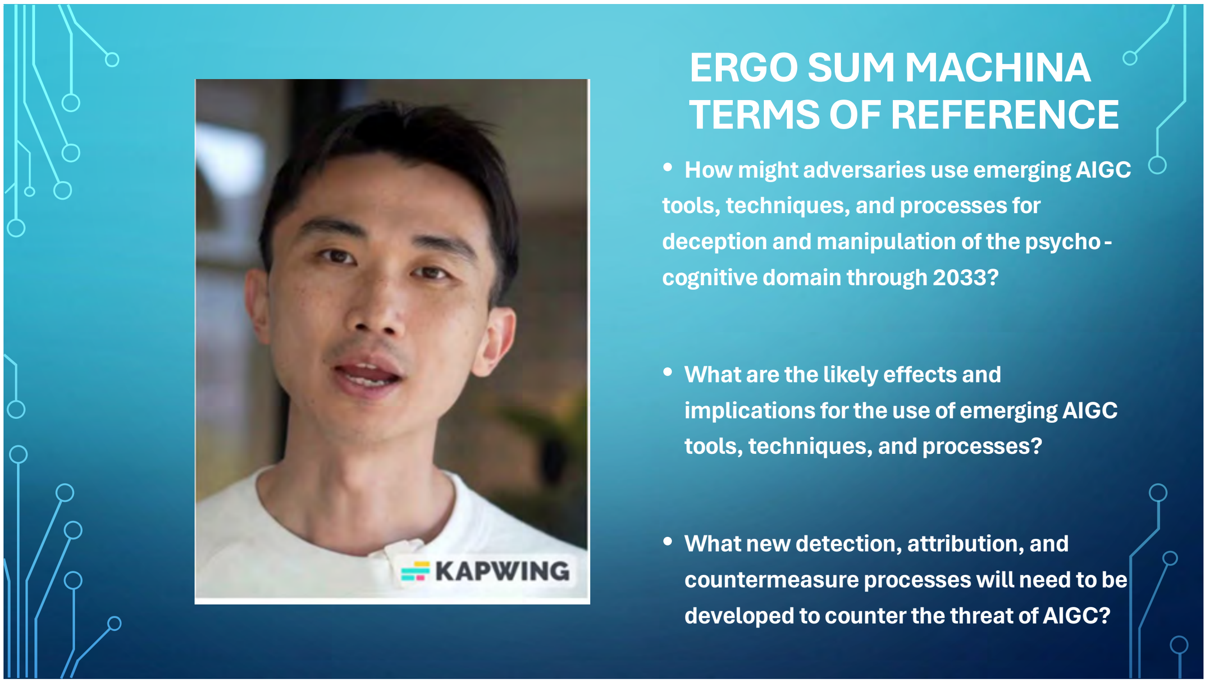
Data produced by existing AI models to generate the massive volume of high-quality, unbiased, cheap data needed for training other AI models.

Psycho-Cognitive Domain

“influence the will of a service member”

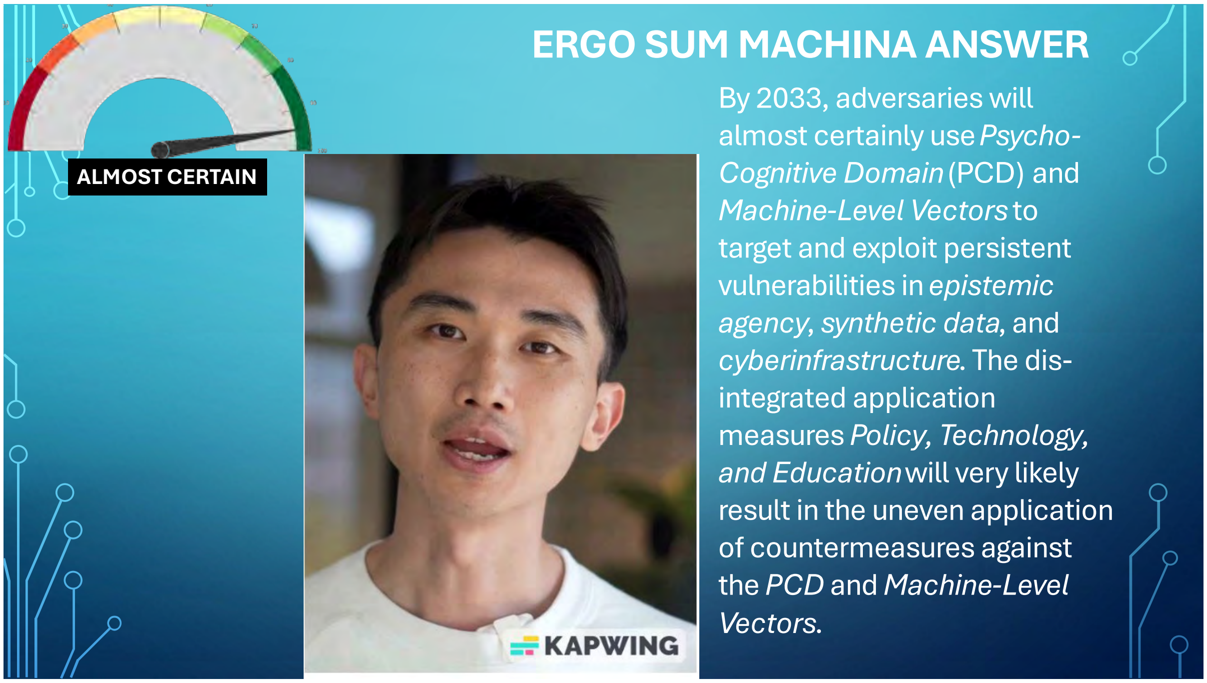
ANALYTICAL CONFIDENCE

Our overall estimate is **MODERATE**



ERGO SUM MACHINA TERMS OF REFERENCE

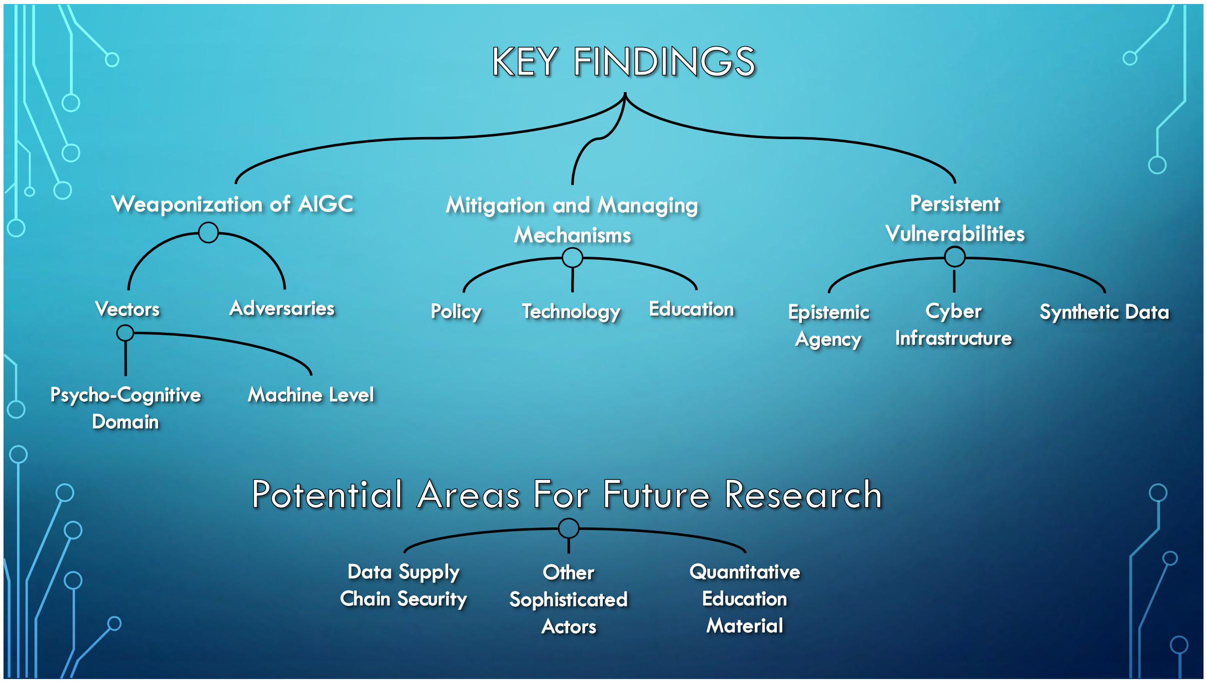
- How might adversaries use emerging AIGC tools, techniques, and processes for deception and manipulation of the psycho-cognitive domain through 2033?
- What are the likely effects and implications for the use of emerging AIGC tools, techniques, and processes?
- What new detection, attribution, and countermeasure processes will need to be developed to counter the threat of AIGC?



ERGO SUM MACHINA ANSWER

ALMOST CERTAIN

By 2033, adversaries will almost certainly use *Psycho-Cognitive Domain* (PCD) and *Machine-Level Vectors* to target and exploit persistent vulnerabilities in *epistemic agency*, *synthetic data*, and *cyberinfrastructure*. The dis-integrated application measures *Policy, Technology, and Education* will very likely result in the uneven application of countermeasures against the *PCD* and *Machine-Level Vectors.*



KEY FINDINGS

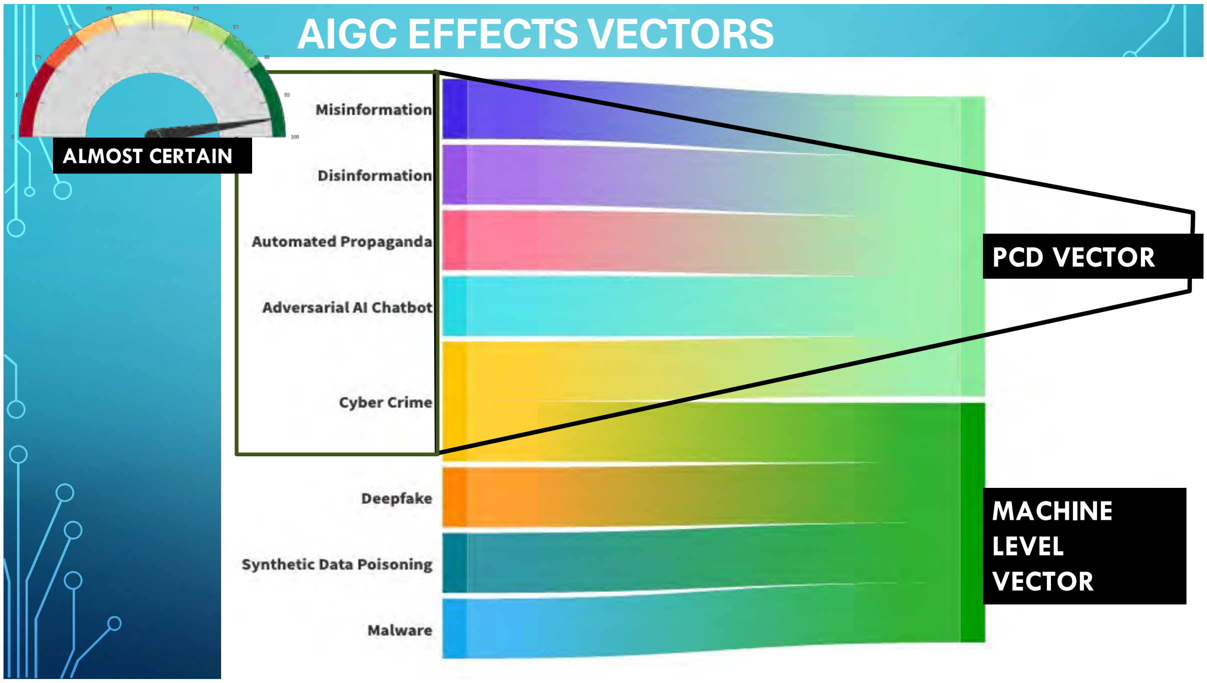
- Weaponization of AIGC
  - Vectors
    - Psycho-Cognitive Domain
    - Machine Level
  - Adversaries
- Mitigation and Managing Mechanisms
  - Policy
  - Technology
  - Education
- Persistent Vulnerabilities
  - Epistemic Agency
  - Cyber Infrastructure
  - Synthetic Data

Potential Areas For Future Research

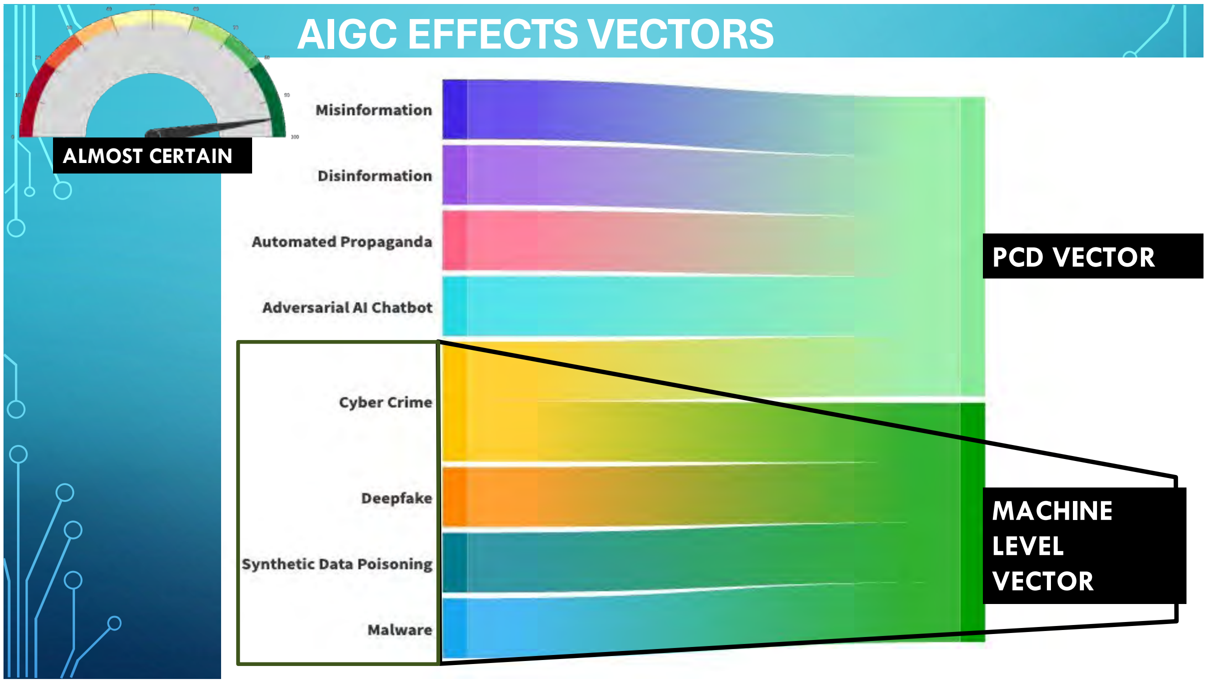
- Data Supply Chain Security
- Other Sophisticated Actors
- Quantitative Education Material

KEY FINDING ONE

Weaponization of AIGC



AIGC EFFECTS VECTORS



AIGC EFFECTS VECTORS

ALMOST CERTAIN

Misinformation

Disinformation

Automated Propaganda

Adversarial AI Chatbot

Cyber Crime

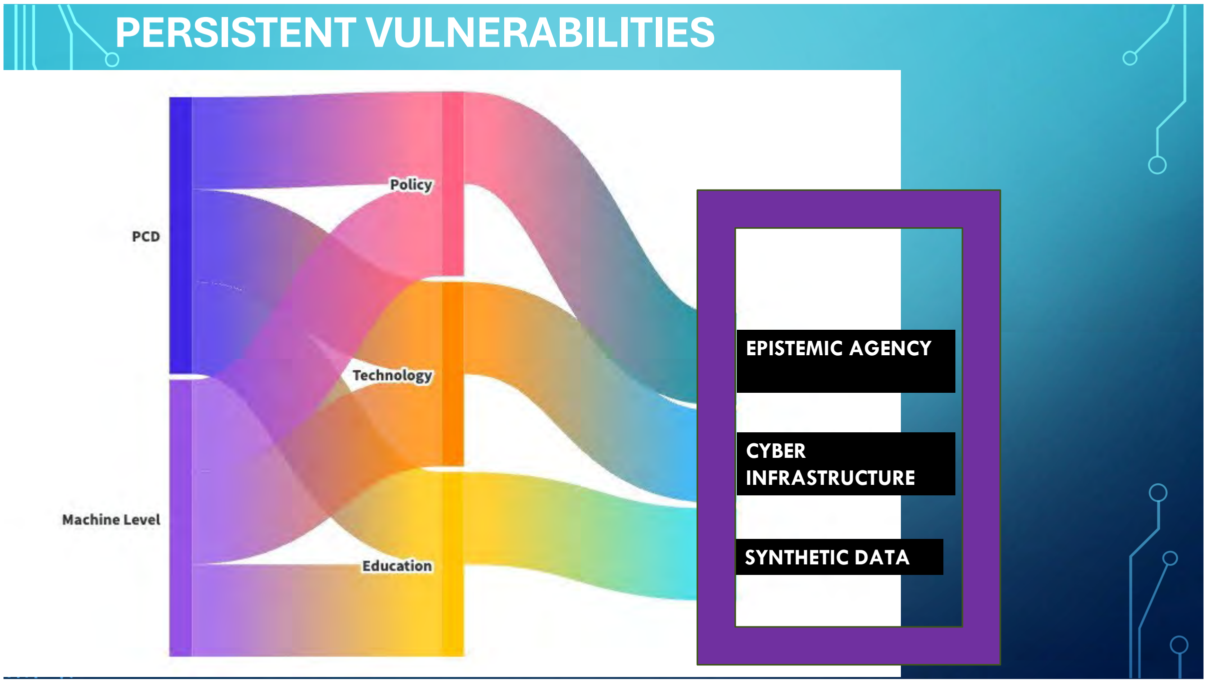
Deepfake

Synthetic Data Poisoning

Malware

PCD VECTOR

MACHINE LEVEL VECTOR



PERSISTENT VULNERABILITIES

PCD

Machine Level

Policy

Technology

Education

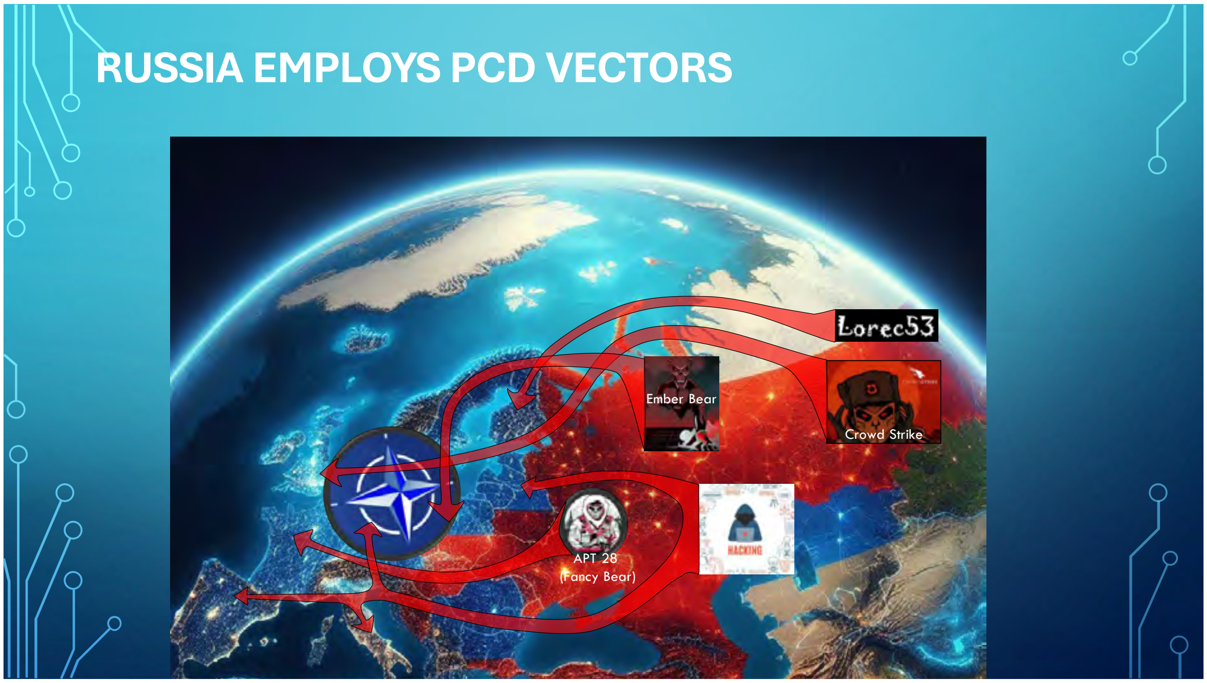
EPISTEMIC AGENCY

CYBER INFRASTRUCTURE

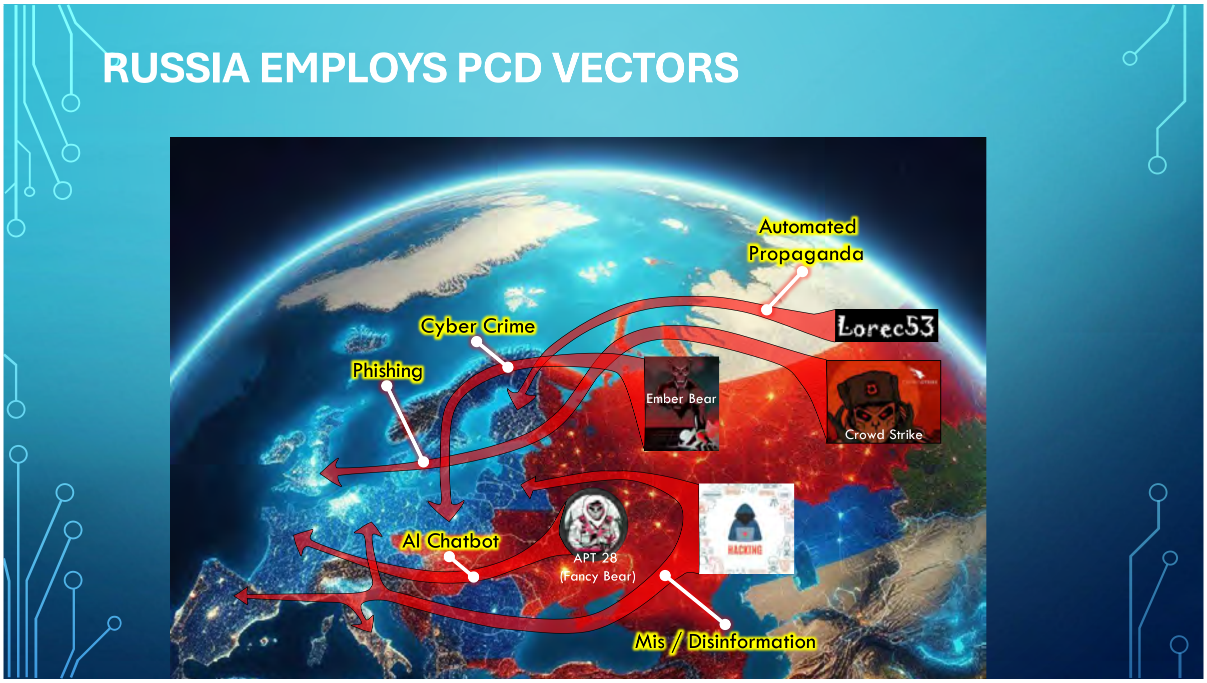
SYNTHETIC DATA

ADVERSARIES & CAPABILITIES

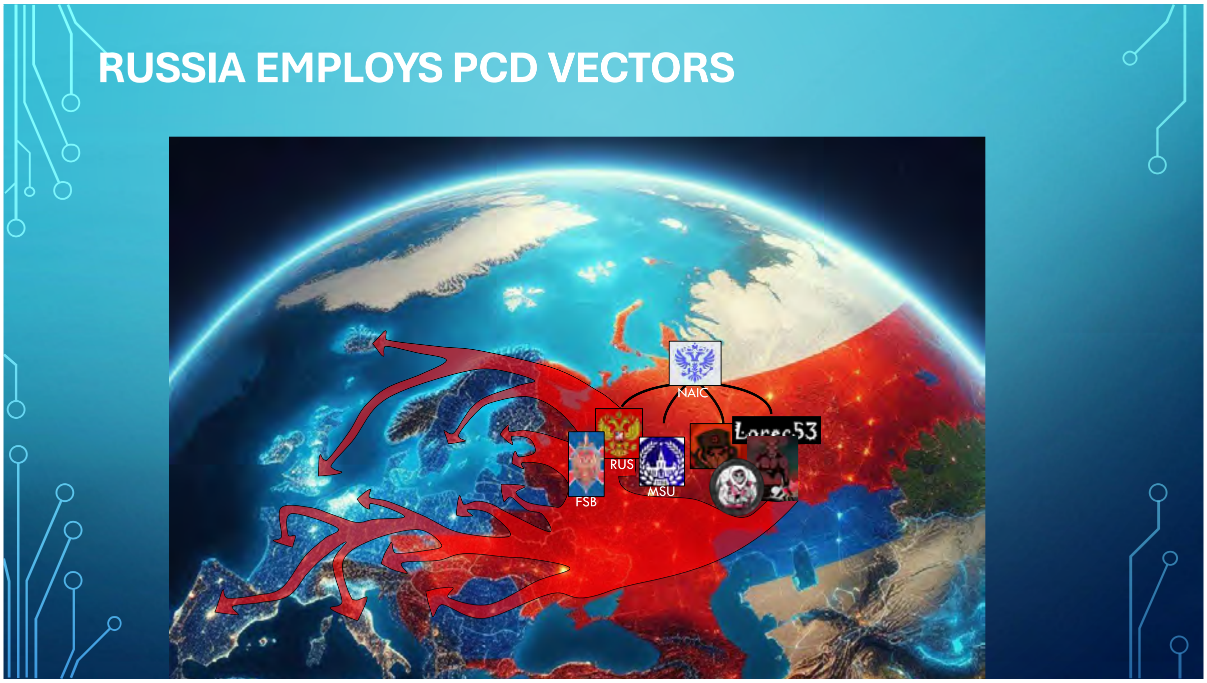
Powered by DALL-E 3



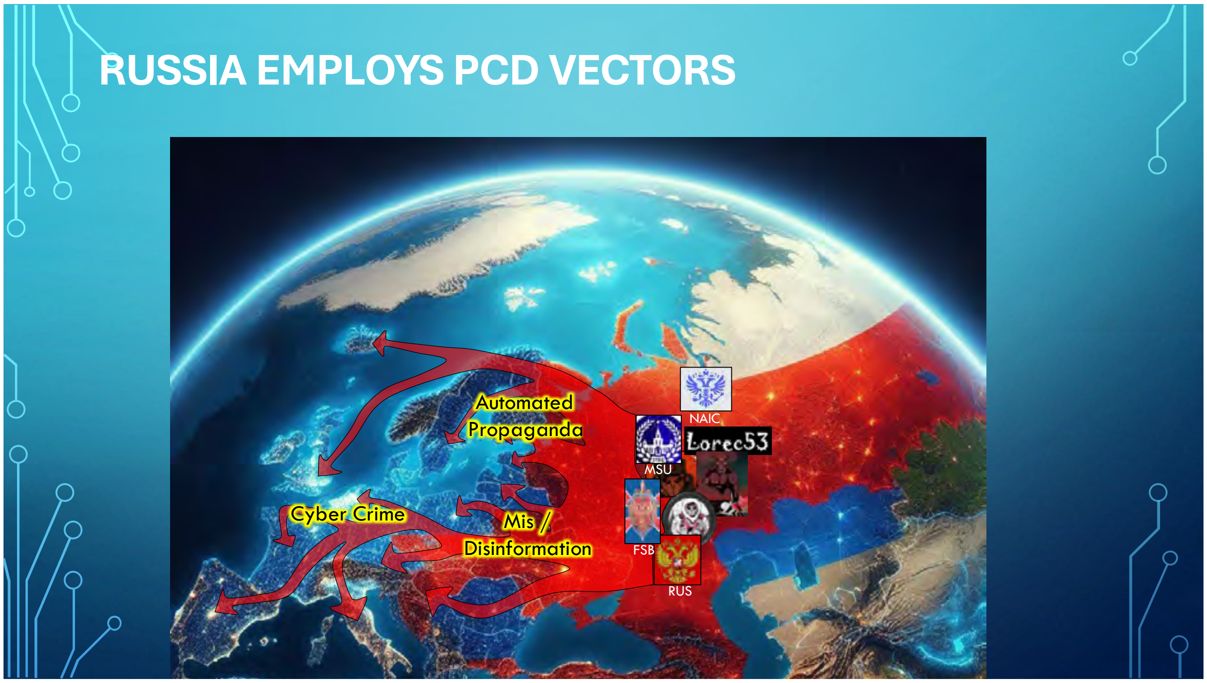
RUSSIA EMPLOYS PCD VECTORS



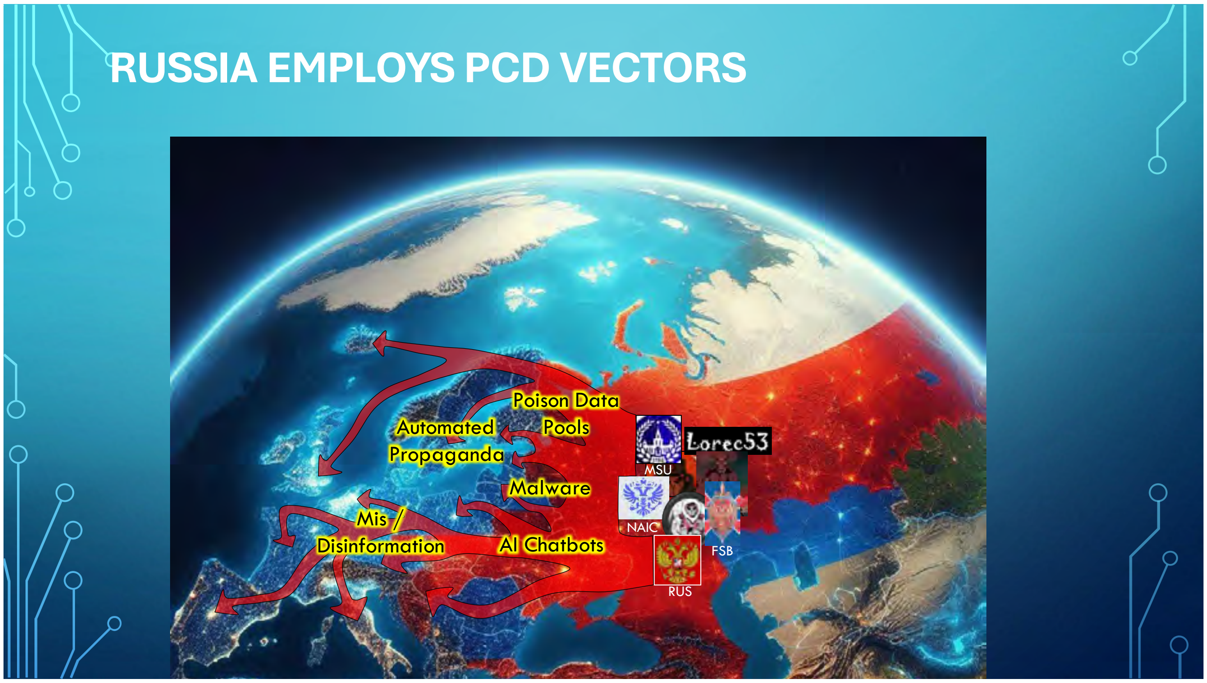
RUSSIA EMPLOYS PCD VECTORS



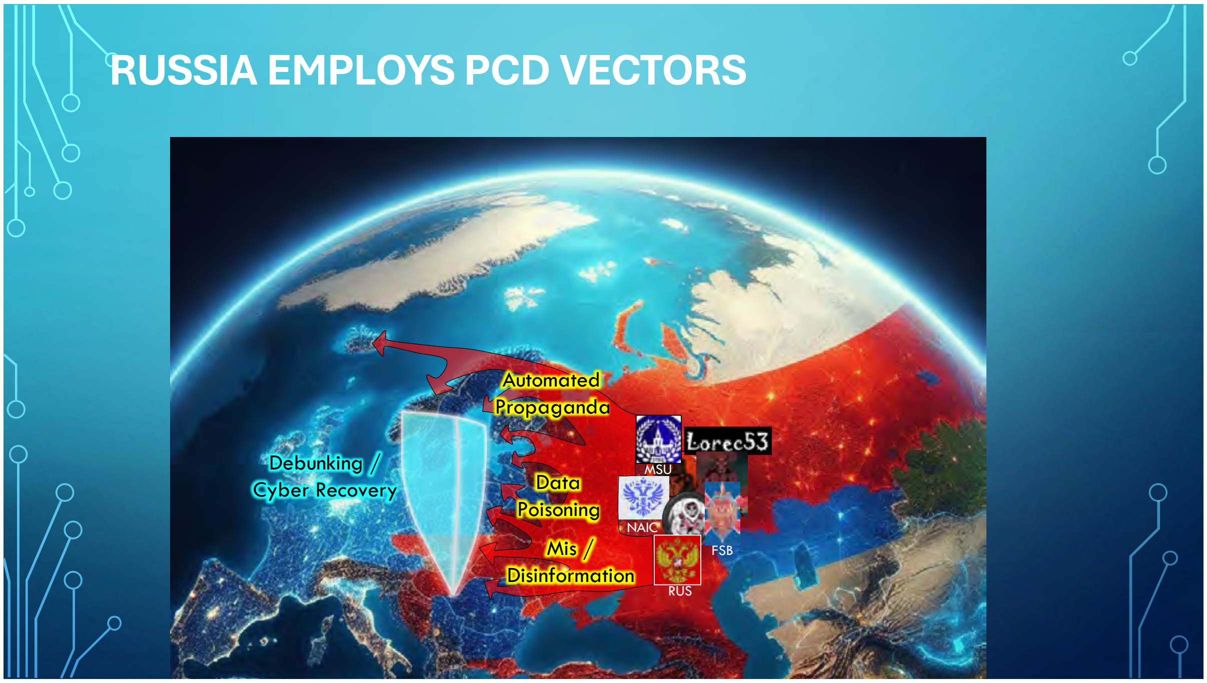
RUSSIA EMPLOYS PCD VECTORS



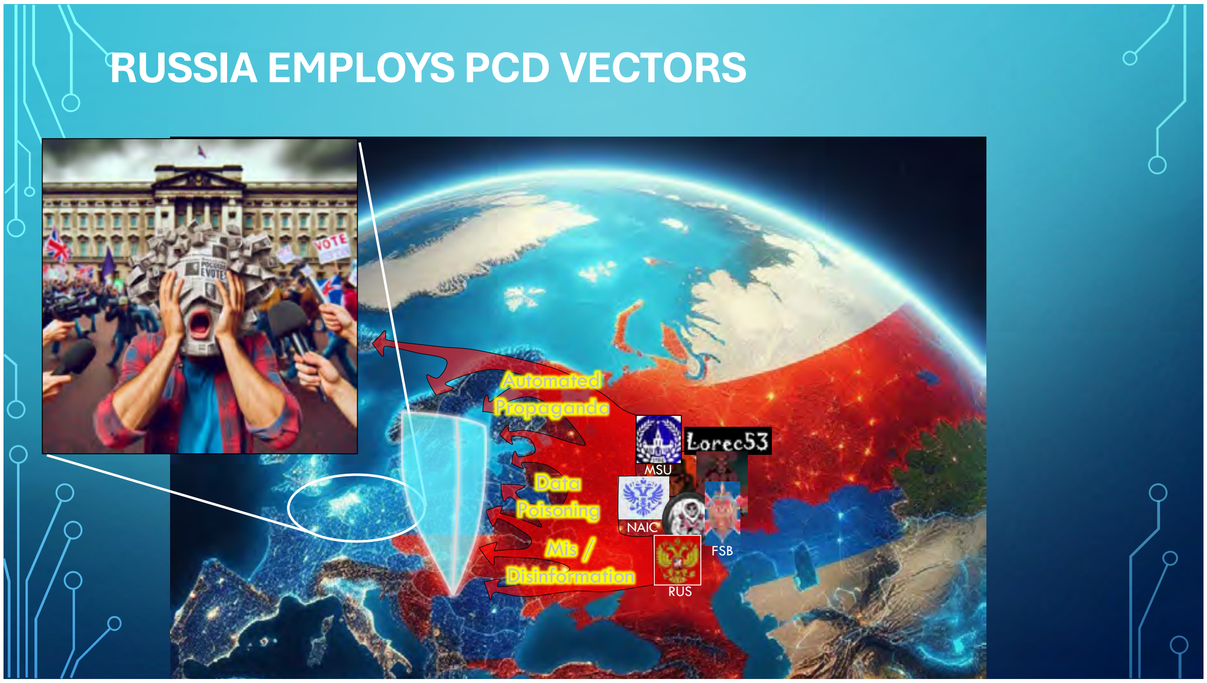
RUSSIA EMPLOYS PCD VECTORS



RUSSIA EMPLOYS PCD VECTORS



RUSSIA EMPLOYS PCD VECTORS



RUSSIA EMPLOYS PCD VECTORS

CHINA EMPLOYS MACHINE LEVEL VECTORS

CHINA EMPLOYS MACHINE LEVEL VECTORS

CHINA EMPLOYS MACHINE LEVEL VECTORS

Cyberspace
Administration
of China

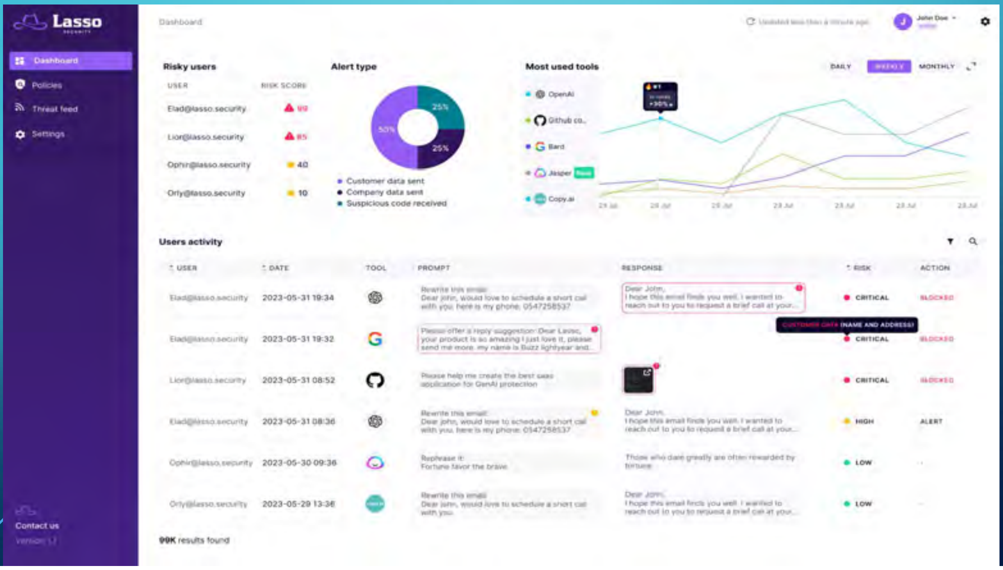
CHINA EMPLOYS MACHINE LEVEL VECTORS

Cyber Crime
Deep Fake
Malware
Synthetic Data
Poisoning

CHINA EMPLOYS MACHINE LEVEL VECTORS

CHINA EMPLOYS MACHINE LEVEL VECTORS

CHINA EMPLOYS MACHINE LEVEL VECTORS



Risky users

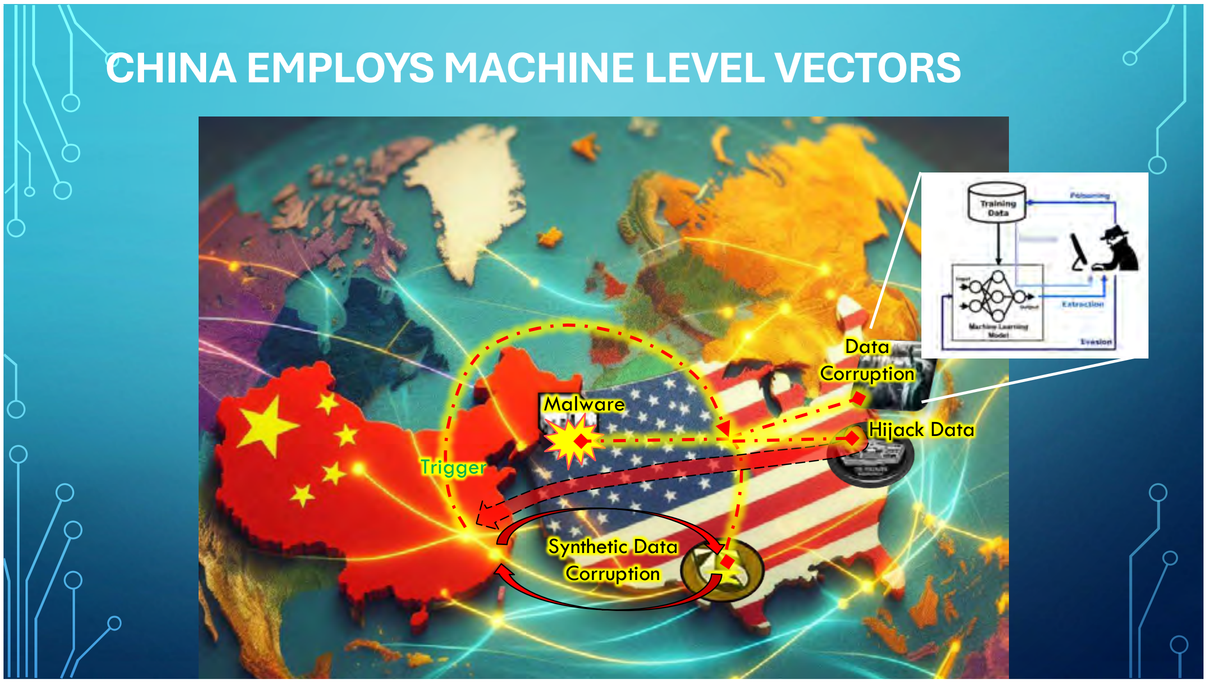
USER	RISK SCORE
Elad@lasso.security	99
Lior@lasso.security	85
Ophir@lasso.security	40
Orly@lasso.security	10

Users activity

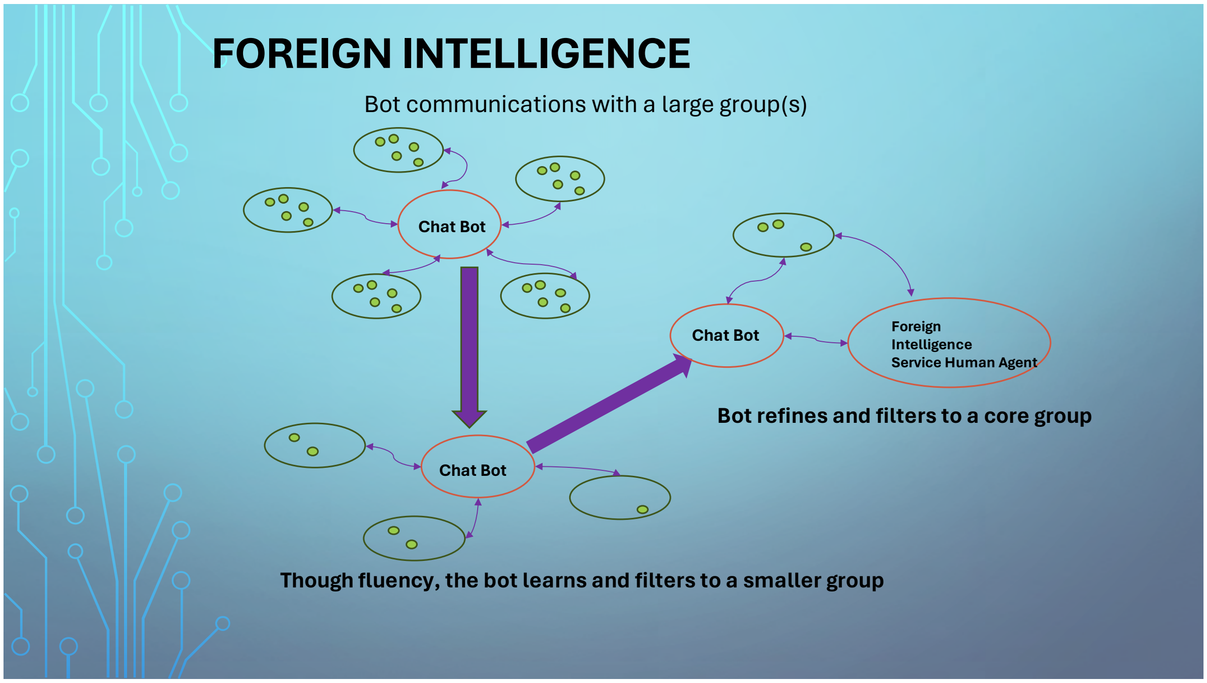
USER	DATE	TOOL	PROMPT	RESPONSE	RISK	ACTION
Elad@lasso.security	2023-05-31 19:34		Rewrite this email: Dear john, would love to schedule a short call with you. here is my phone: 0547258537	Dear John, I hope this email finds you well. I wanted to reach out to you to request a brief call at your...	CRITICAL	BLOCKED
Elad@lasso.security	2023-05-31 19:32	G	Please offer a reply suggestion: Dear Lasso, your product is so amazing I just love it, please send me more. my name is Buzz lightyear and...		CUSTOMER DATA (NAME AND ADDRESS) CRITICAL	BLOCKED
Lior@lasso.security	2023-05-31 08:52		Please help me create the best saas application for GenAI protection		CRITICAL	BLOCKED
Elad@lasso.security	2023-05-31 08:36		Rewrite this email: Dear john, would love to schedule a short call with you. here is my phone: 0547258537	Dear John, I hope this email finds you well. I wanted to reach out to you to request a brief call at your...	HIGH	ALERT
Ophir@lasso.security	2023-05-30 09:36		Rephrase it: Fortune favor the brave	Those who dare greatly are often rewarded by fortune	LOW	-
Orly@lasso.security	2023-05-29 13:36		Rewrite this email: Dear john, would love to schedule a short call with you	Dear John, I hope this email finds you well. I wanted to reach out to you to request a brief call at your...	LOW	-

99K results found

CHINA EMPLOYS MACHINE LEVEL VECTORS



CHINA EMPLOYS MACHINE LEVEL VECTORS



FOREIGN INTELLIGENCE

Bot communications with a large group(s)

Chat Bot

Chat Bot

Foreign Intelligence Service Human Agent

Bot refines and filters to a core group

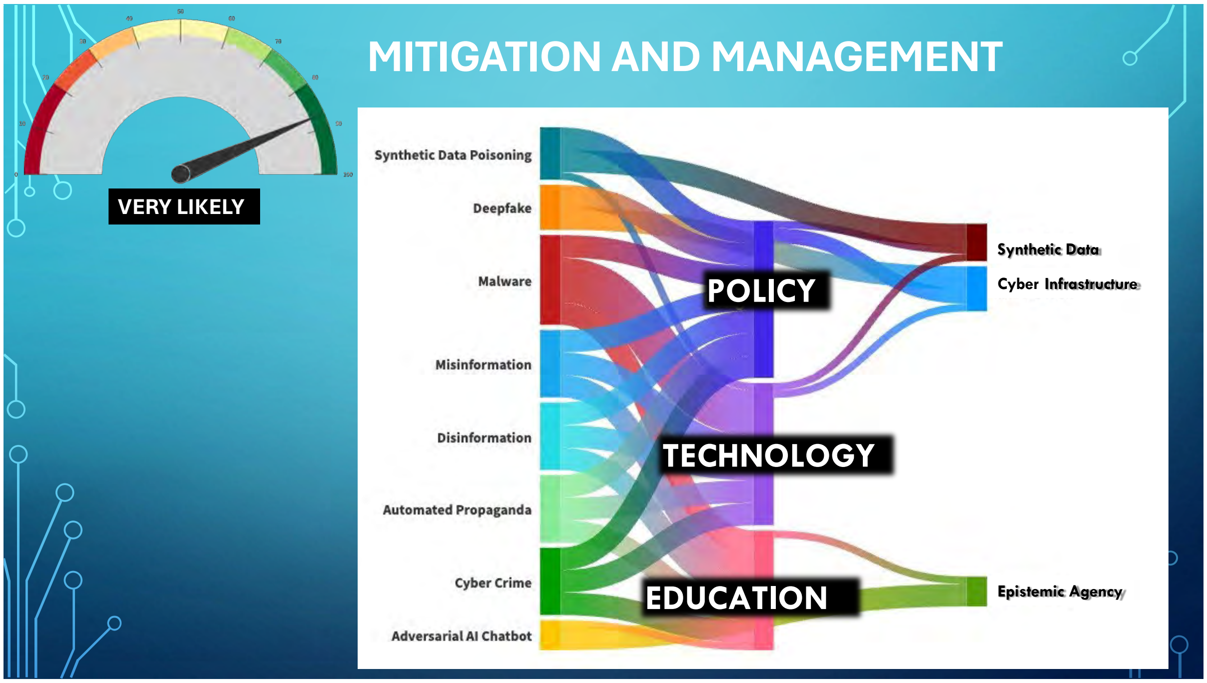
Chat Bot

Though fluency, the bot learns and filters to a smaller group

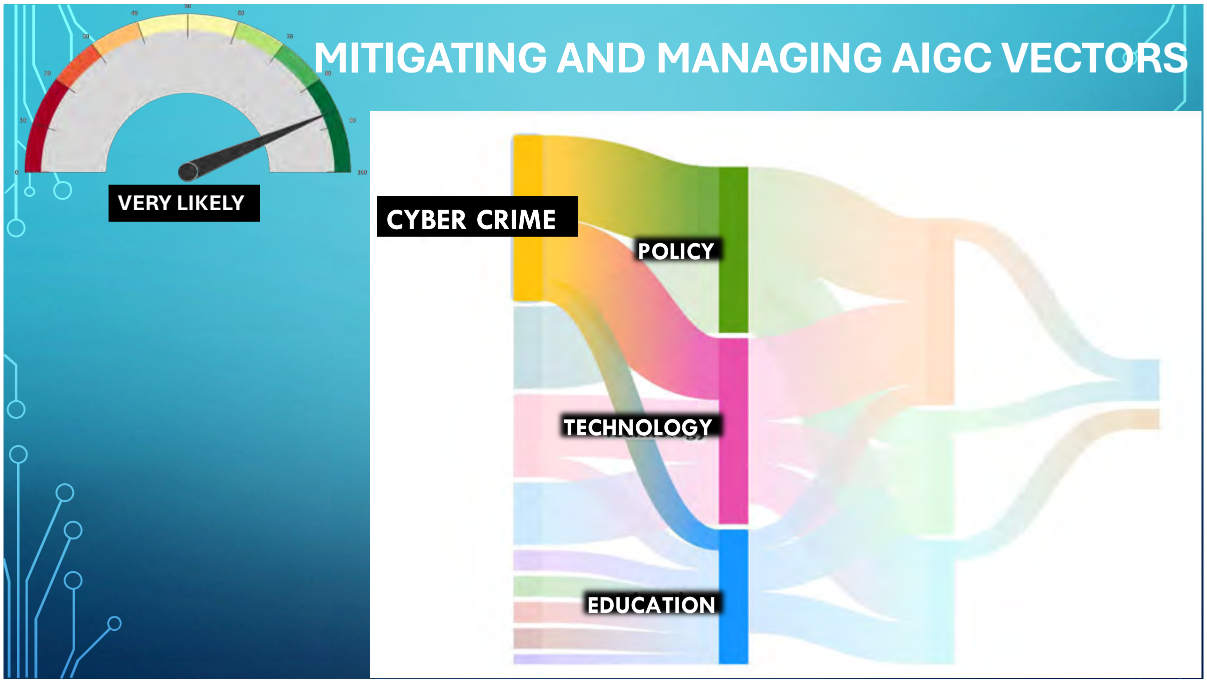
FUTURE OF FOREIGN INTELLIGENCE

KEY FINDING TWO

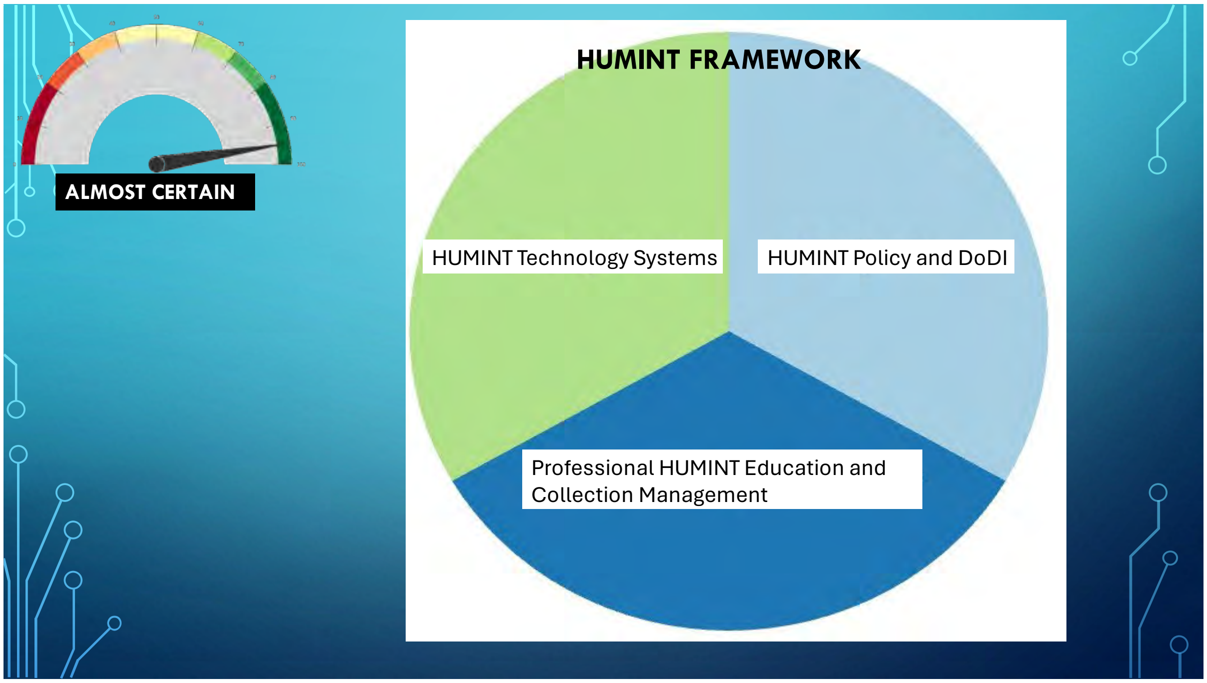
Mitigation/Management Measures



MITIGATION AND MANAGEMENT



MITIGATING AND MANAGING AIGC VECTORS



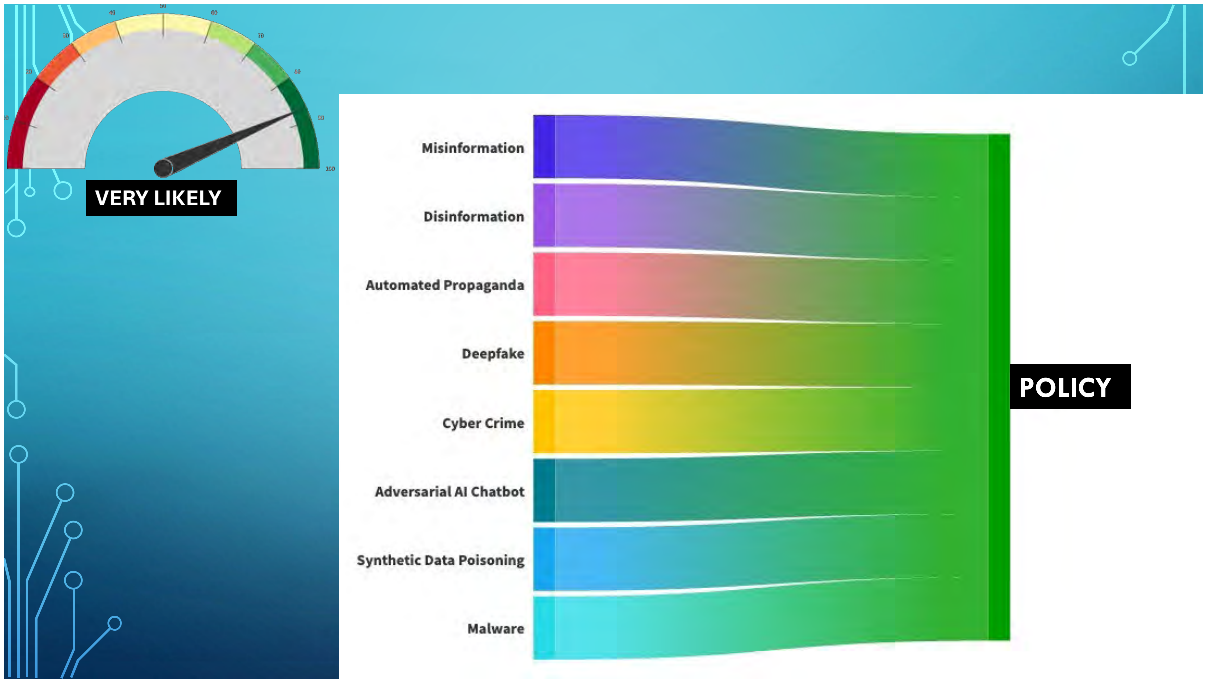
ALMOST CERTAIN

HUMINT FRAMEWORK

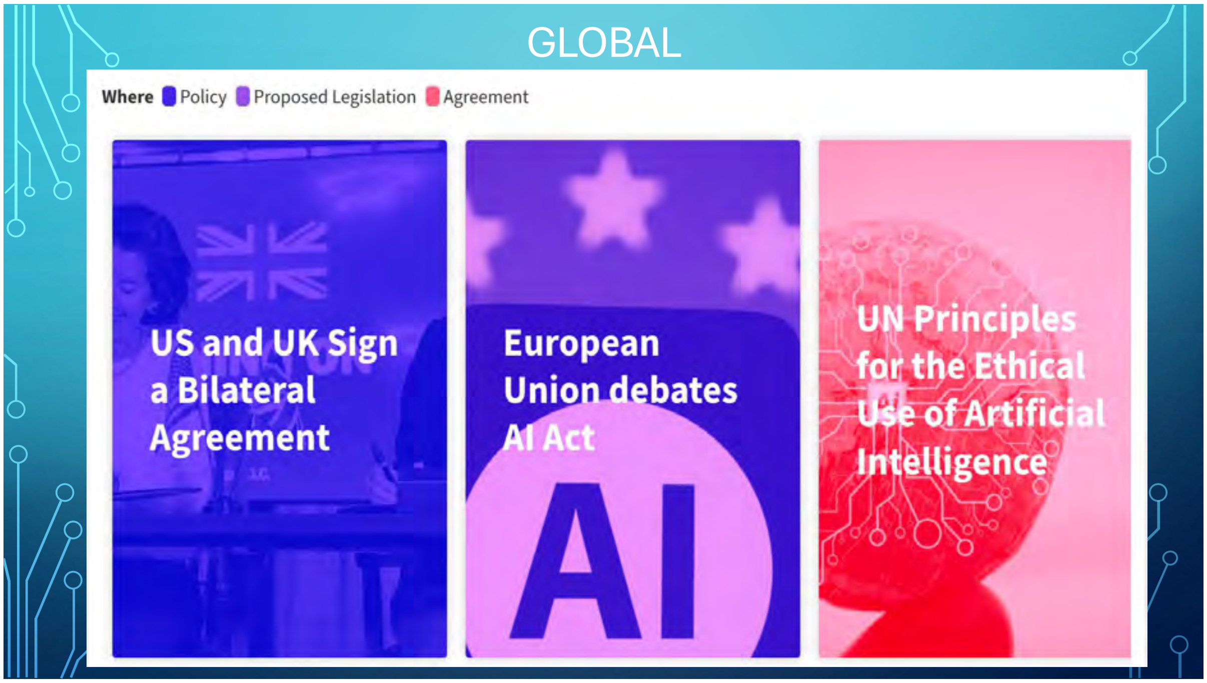
HUMINT Technology Systems

HUMINT Policy and DoDI

Professional HUMINT Education and Collection Management



VERY LIKELY



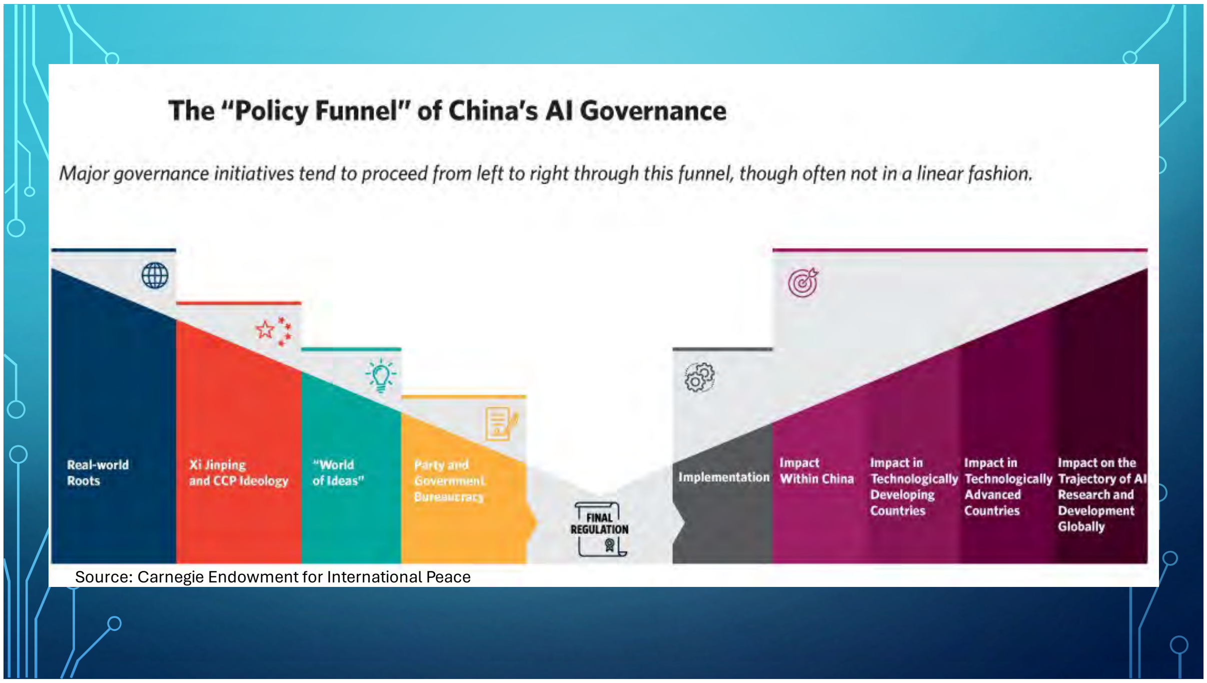
GLOBAL

Where 🟦 Policy 🟪 Proposed Legislation 🟥 Agreement

US and UK Sign a Bilateral Agreement

European Union debates AI Act

UN Principles for the Ethical Use of Artificial Intelligence



The "Policy Funnel" of China's AI Governance

Major governance initiatives tend to proceed from left to right through this funnel, though often not in a linear fashion.

Real-world Roots

Xi Jinping and CCP Ideology

"World of Ideas"

Party and Government Bureaucracy

FINAL REGULATION

Implementation

Impact Within China

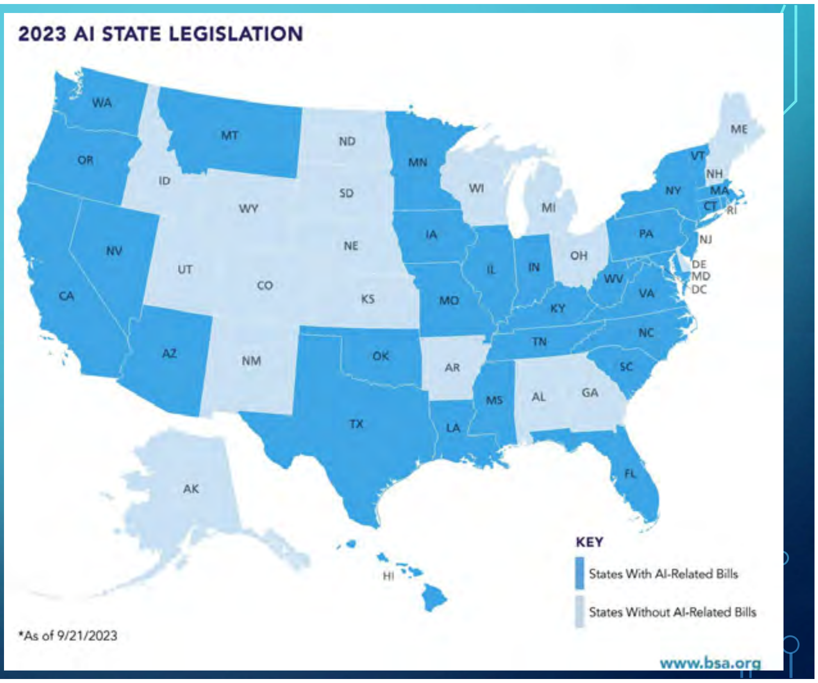
Impact in Technologically Developing Countries

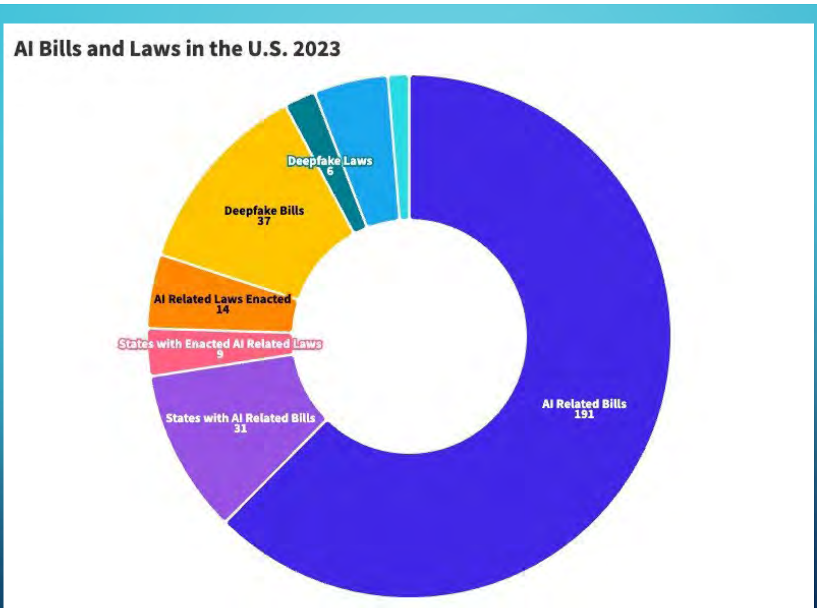
Impact in Technologically Advanced Countries

Impact on the Trajectory of AI Research and Development Globally

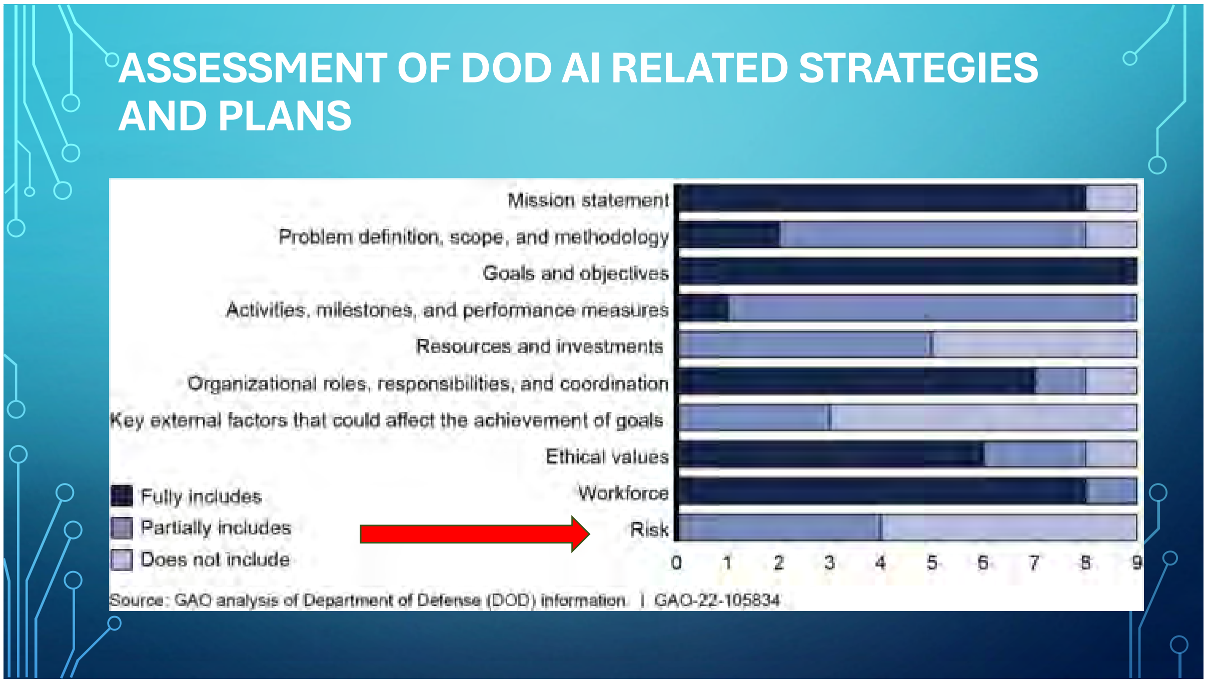
Source: Carnegie Endowment for International Peace

DOMESTIC



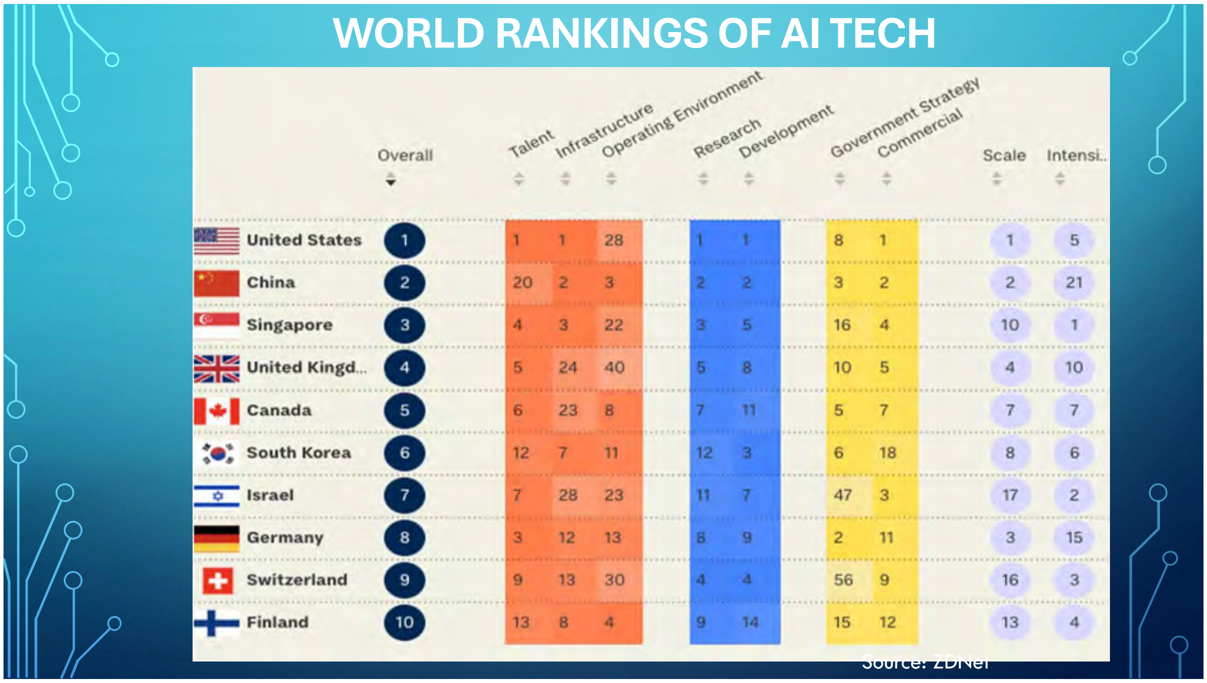


In 2023 AI Bill Submission Increased 441%



ASSESSMENT OF DOD AI RELATED STRATEGIES AND PLANS

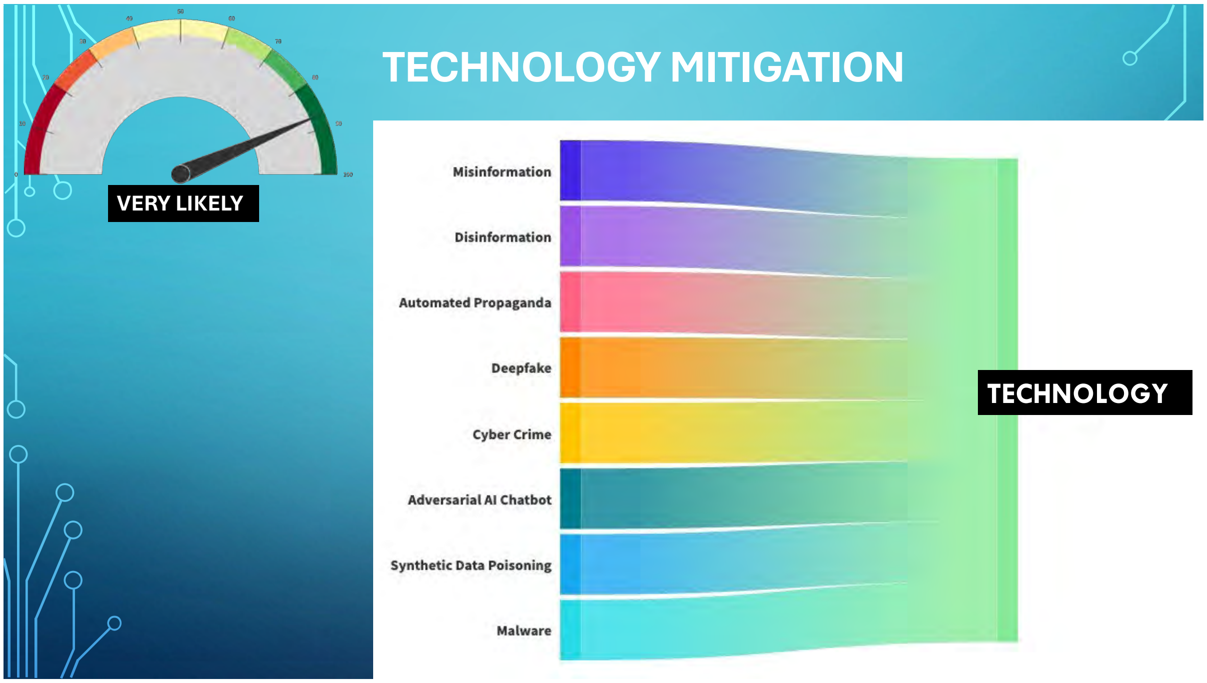
Source: GAO analysis of Department of Defense (DOD) information | GAO-22-105834



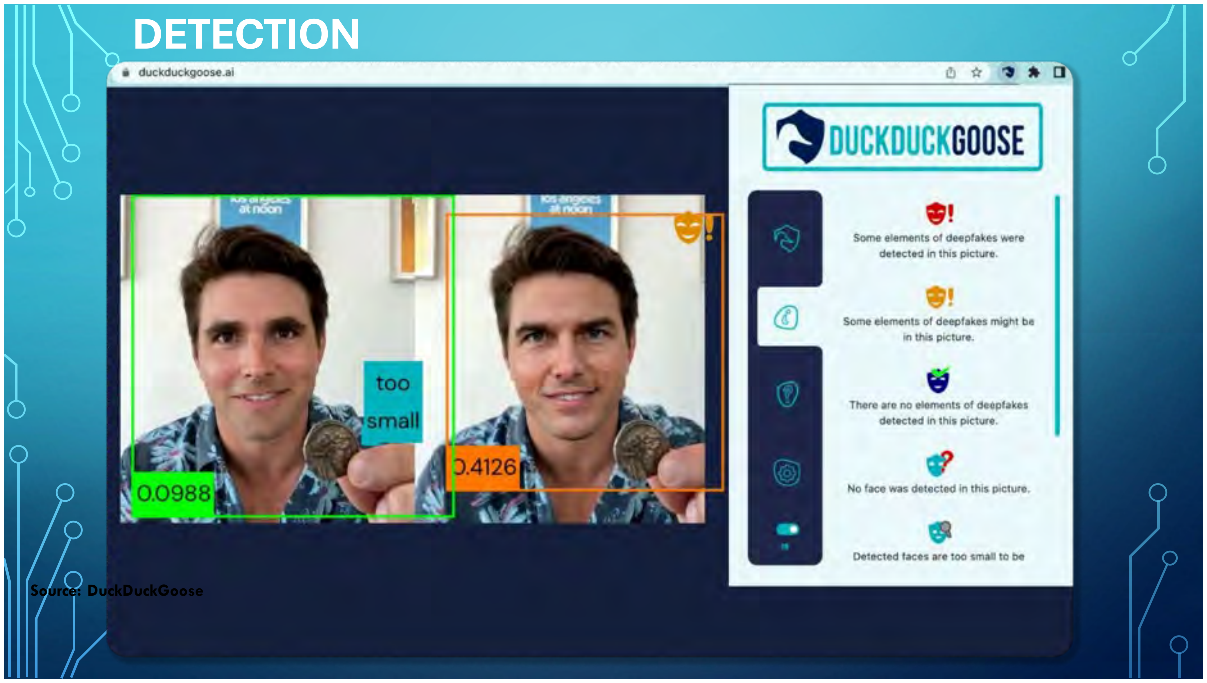
WORLD RANKINGS OF AI TECH

	Overall	Talent	Infrastructure	Operating Environment	Research	Development	Government Strategy	Commercial	Scale	Intensi..
United States	1	1	1	28	1	1	8	1	1	5
China	2	20	2	3	2	2	3	2	2	21
Singapore	3	4	3	22	3	5	16	4	10	1
United Kingd...	4	5	24	40	5	8	10	5	4	10
Canada	5	6	23	8	7	11	5	7	7	7
South Korea	6	12	7	11	12	3	6	18	8	6
Israel	7	7	28	23	11	7	47	3	17	2
Germany	8	3	12	13	8	9	2	11	3	15
Switzerland	9	9	13	30	4	4	56	9	16	3
Finland	10	13	8	4	9	14	15	12	13	4

Source: ZDNet



TECHNOLOGY MITIGATION



DETECTION

Some elements of deepfakes were detected in this picture.

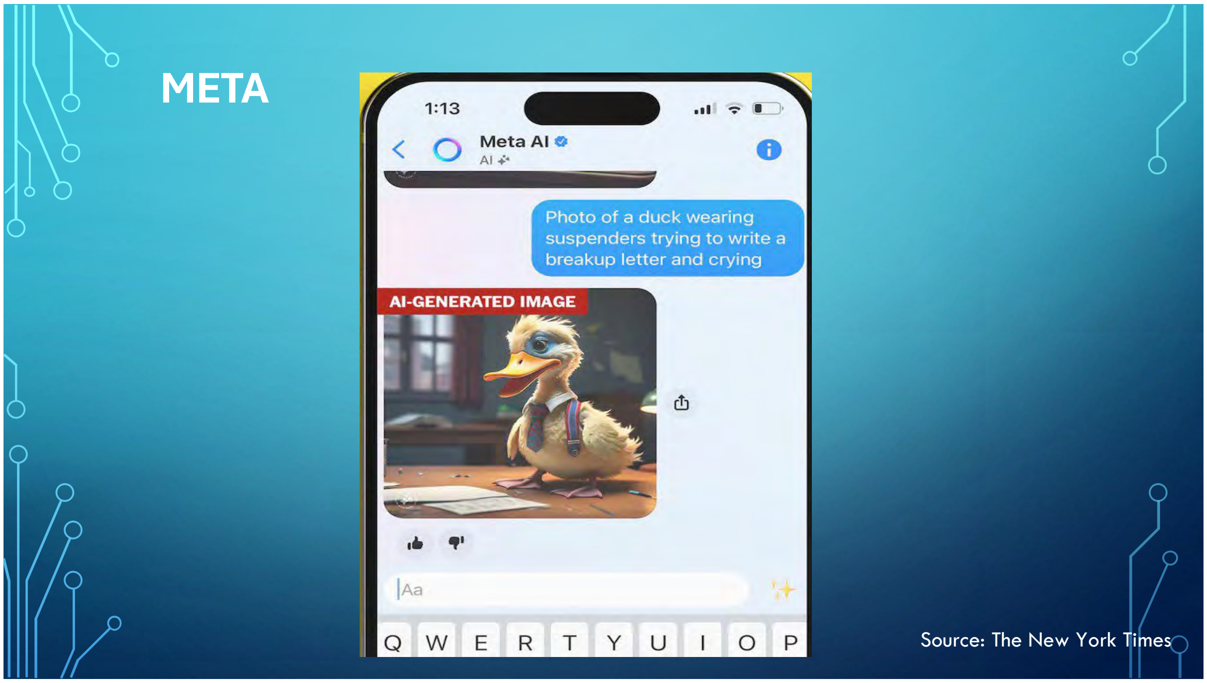
Some elements of deepfakes might be in this picture.

There are no elements of deepfakes detected in this picture.

No face was detected in this picture.

Detected faces are too small to be

Source: DuckDuckGoose

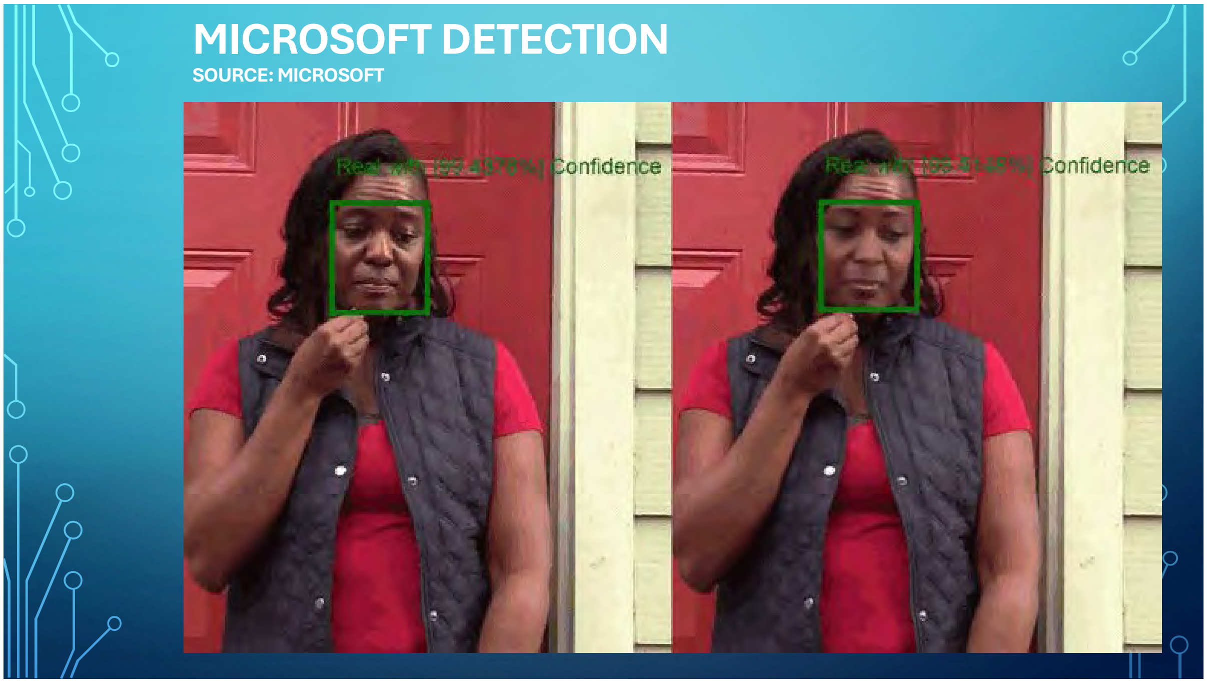


META

Photo of a duck wearing suspenders trying to write a breakup letter and crying

AI-GENERATED IMAGE

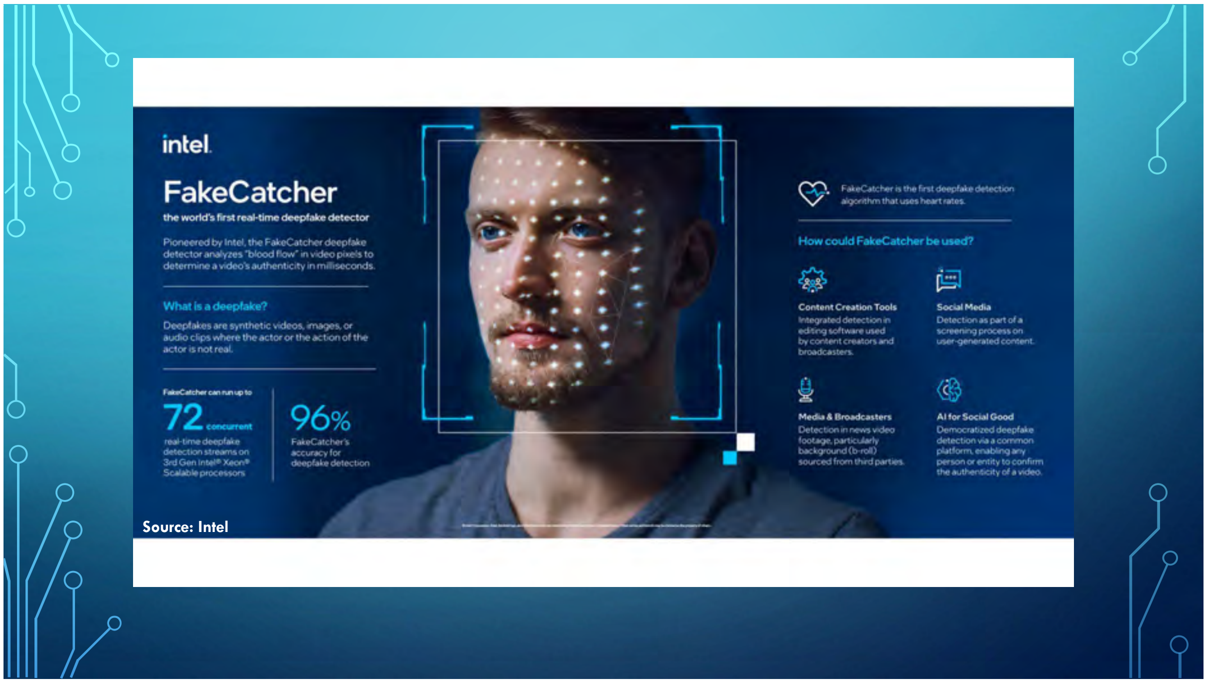
Source: The New York Times



MICROSOFT DETECTION

SOURCE: MICROSOFT

Real with [99.4378%] Confidence
Real with [99.9148%] Confidence



intel.

FakeCatcher

the world's first real-time deepfake detector

Pioneered by Intel, the FakeCatcher deepfake detector analyzes "blood flow" in video pixels to determine a video's authenticity in milliseconds.

What is a deepfake?

Deepfakes are synthetic videos, images, or audio clips where the actor or the action of the actor is not real.

FakeCatcher can run up to

72 concurrent

real-time deepfake detection streams on 3rd Gen Intel® Xeon® Scalable processors

96%

FakeCatcher's accuracy for deepfake detection

FakeCatcher is the first deepfake detection algorithm that uses heart rates.

How could FakeCatcher be used?

Content Creation Tools
Integrated detection in editing software used by content creators and broadcasters.

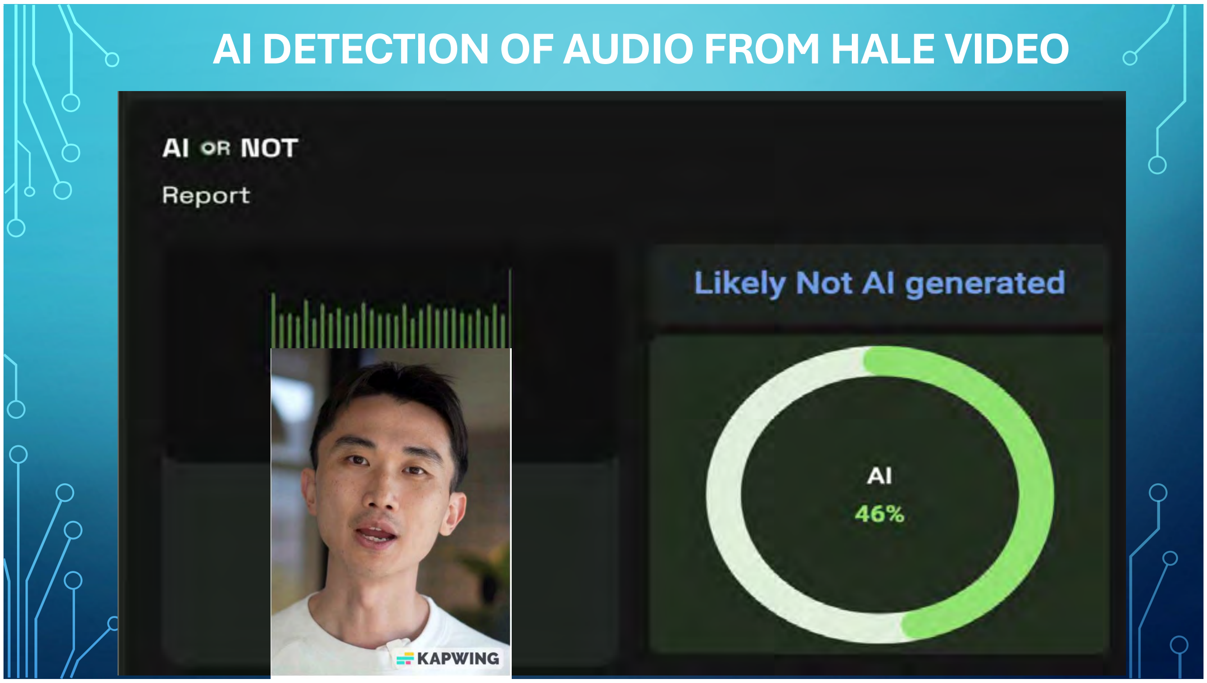
Social Media
Detection as part of a screening process on user-generated content.

Media & Broadcasters
Detection in news video footage, particularly background (b-roll) sourced from third parties.

AI for Social Good
Democratized deepfake detection via a common platform, enabling any person or entity to confirm the authenticity of a video.

Source: Intel

AI DETECTION OF AUDIO FROM HALE VIDEO



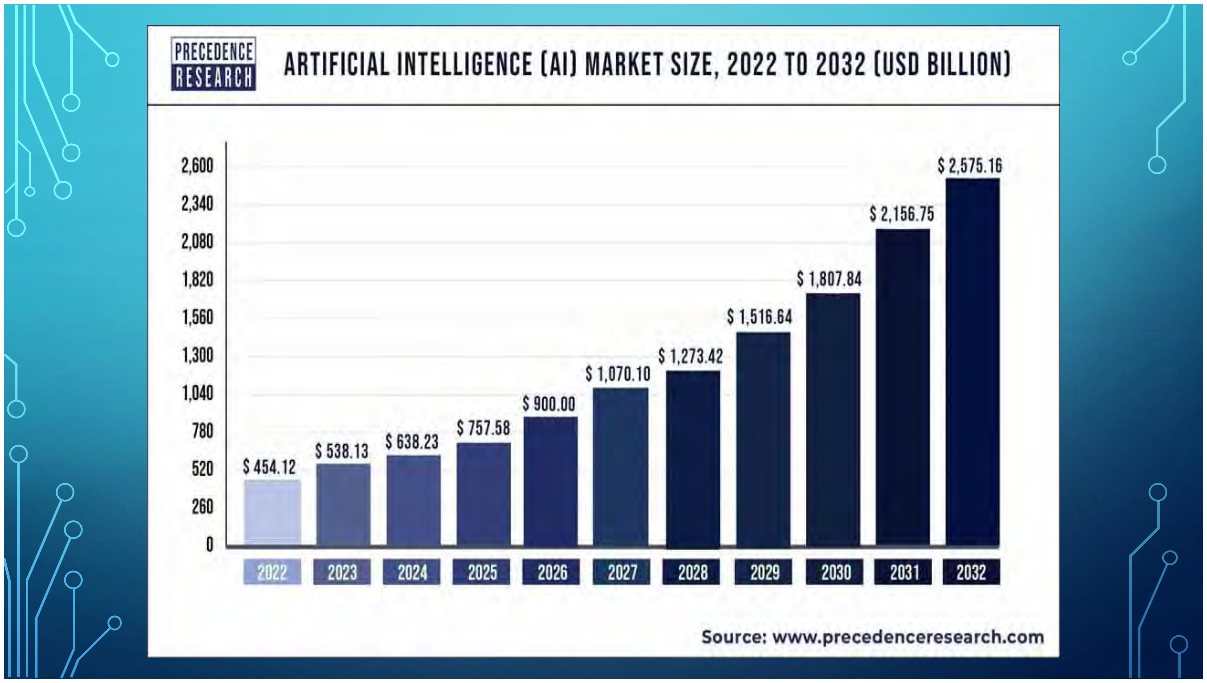
AI OR NOT

Report

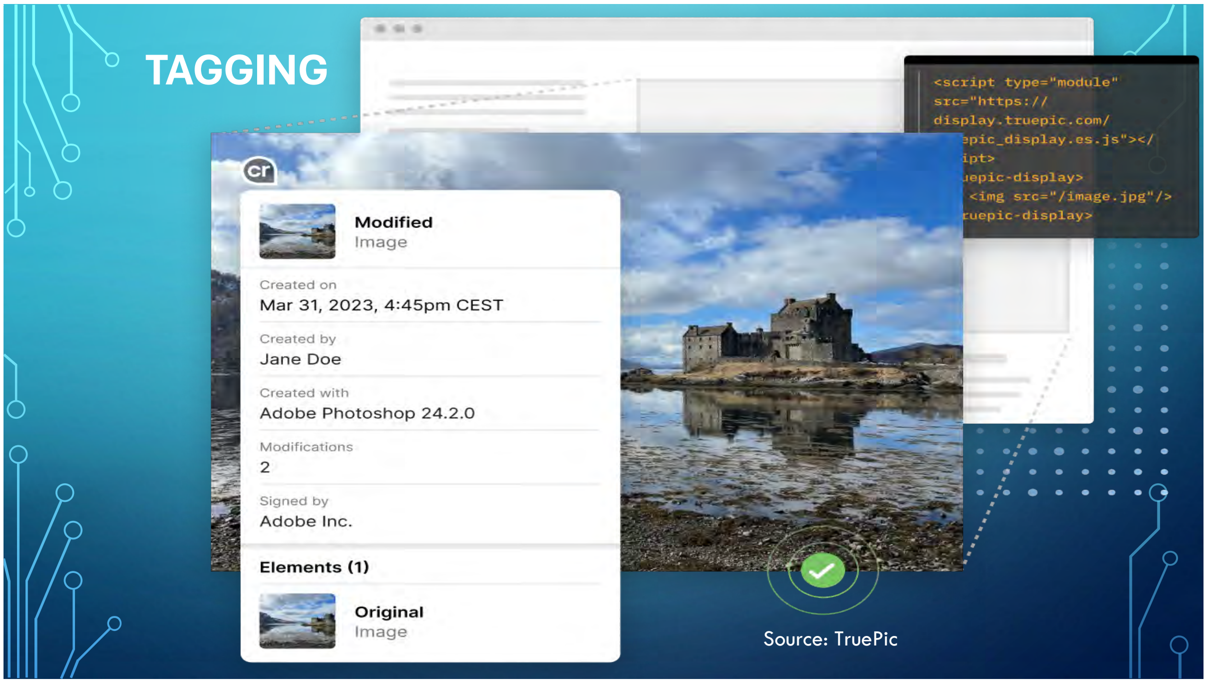
Likely Not AI generated

AI

46%



Source: www.precedenceresearch.com



TAGGING

```
<script type="module"
src="https://
display.truepic.com/
epic_display.es.js"></
ipt>
uepic-display>
 <img src="/image.jpg"/>
ruepic-display>
```

Modified
Image

Created on
Mar 31, 2023, 4:45pm CEST

Created by
Jane Doe

Created with
Adobe Photoshop 24.2.0

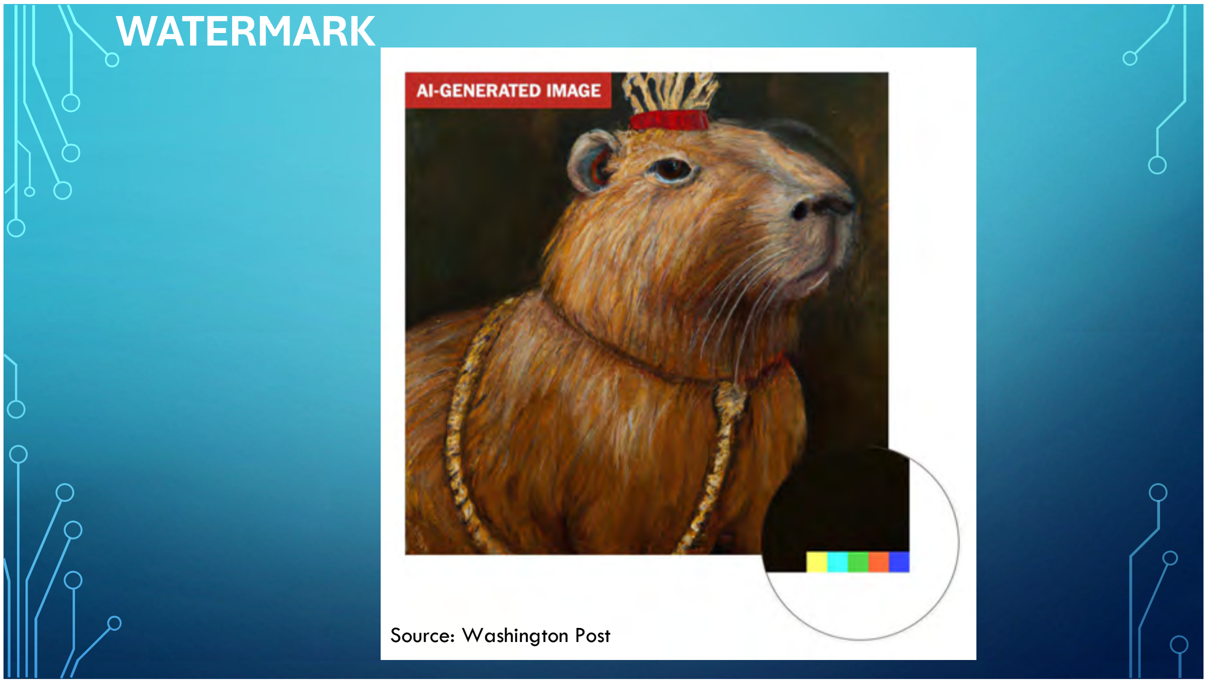
Modifications
2

Signed by
Adobe Inc.

Elements (1)

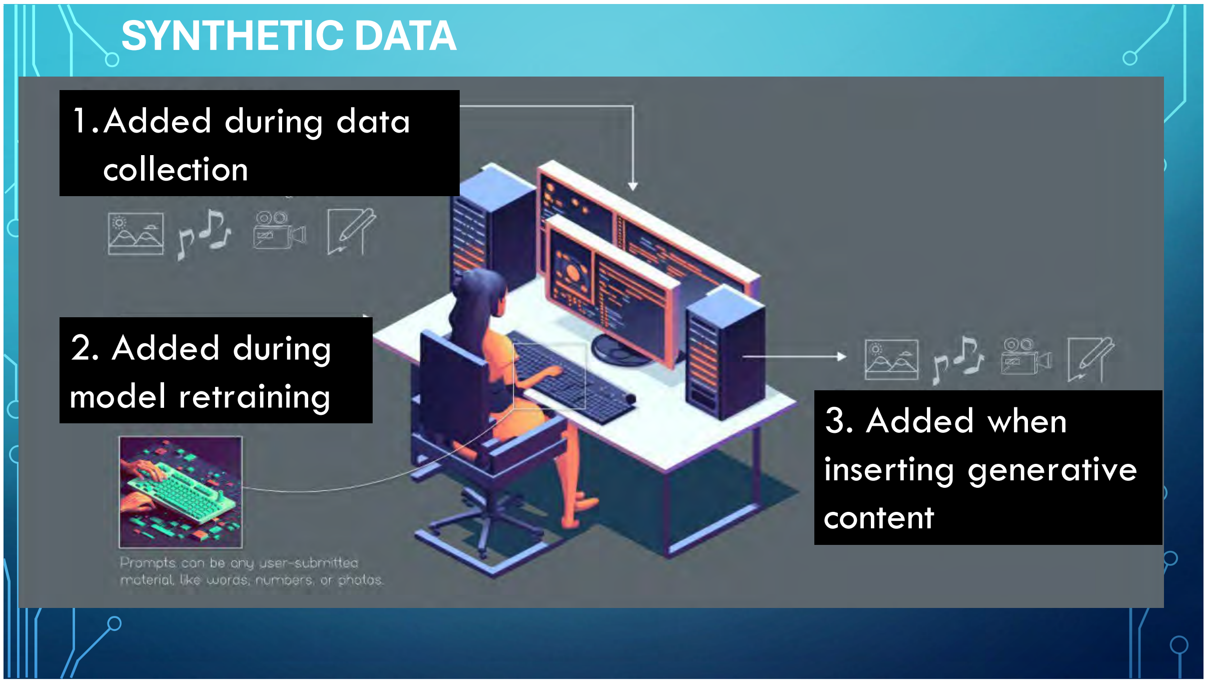
Original
Image

Source: TruePic



WATERMARK

Source: Washington Post



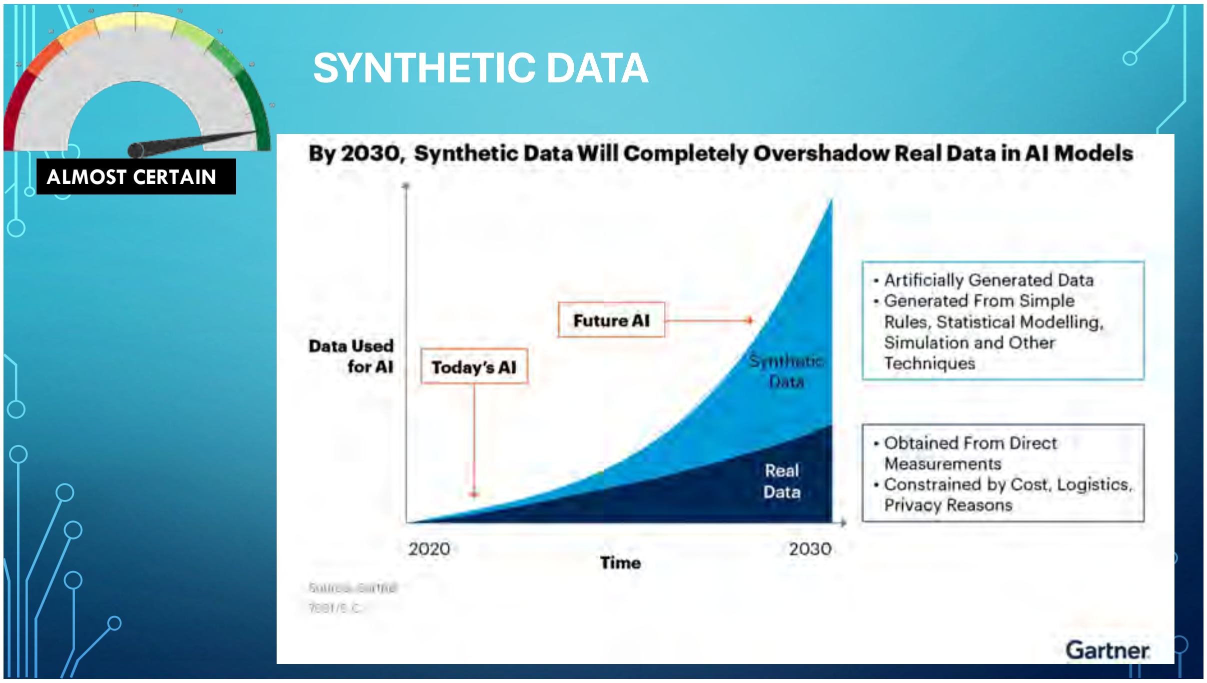
SYNTHETIC DATA

1. Added during data collection

2. Added during model retraining

3. Added when inserting generative content

Prompts can be any user-submitted material, like words, numbers, or photos.



SYNTHETIC DATA

ALMOST CERTAIN

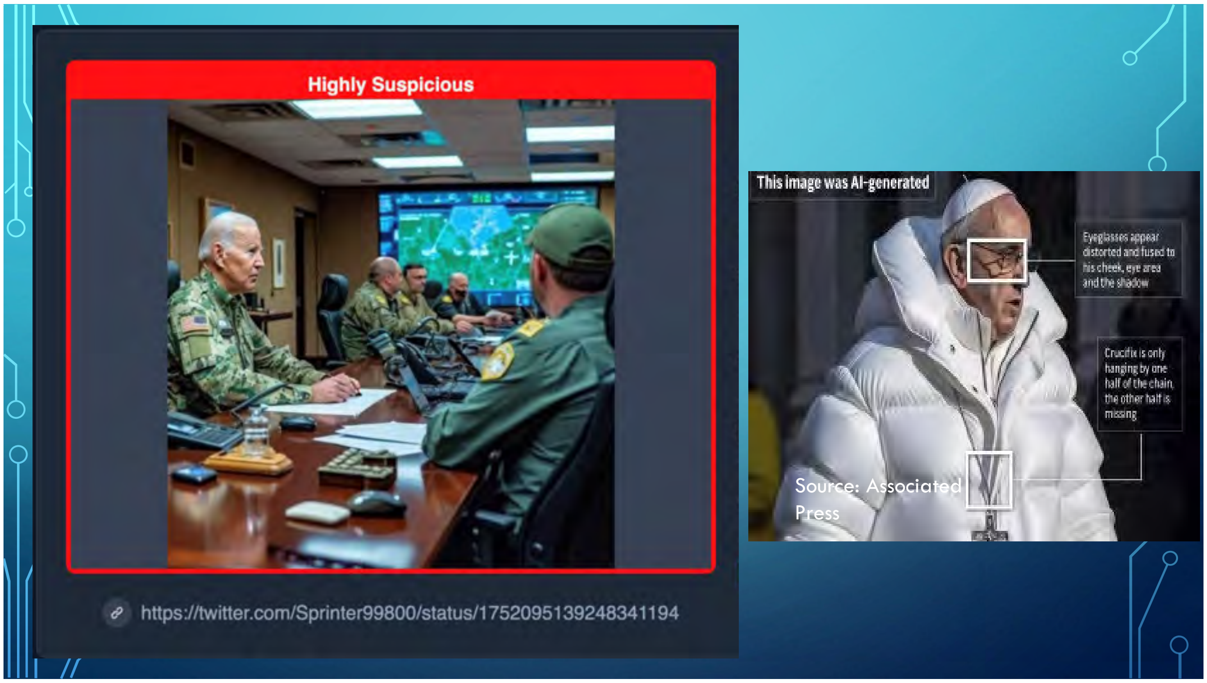
By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models

- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

Source: Gartner
768175_C

Gartner



Highly Suspicious

https://twitter.com/Sprinter99800/status/1752095139248341194

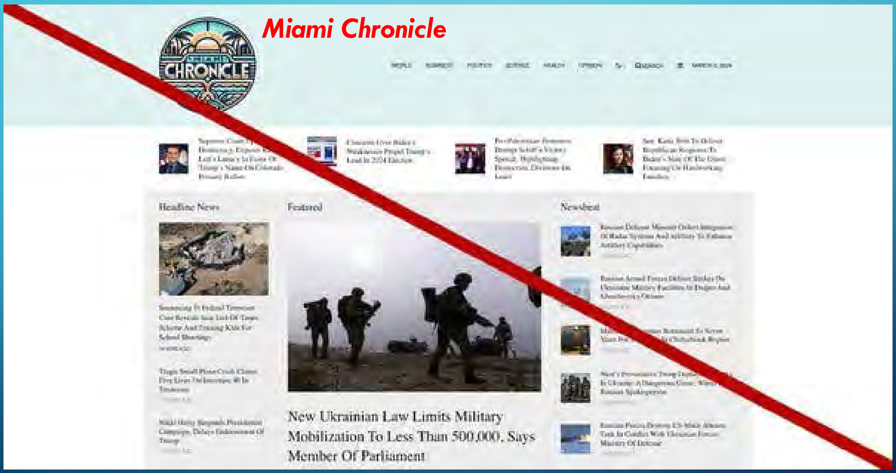
This image was AI-generated

Eyeglasses appear distorted and fused to his cheek, eye area and the shadow

Crucifix is only hanging by one half of the chain, the other half is missing

Source: Associated Press

FAKE NEWS SITE



Miami Chronicle

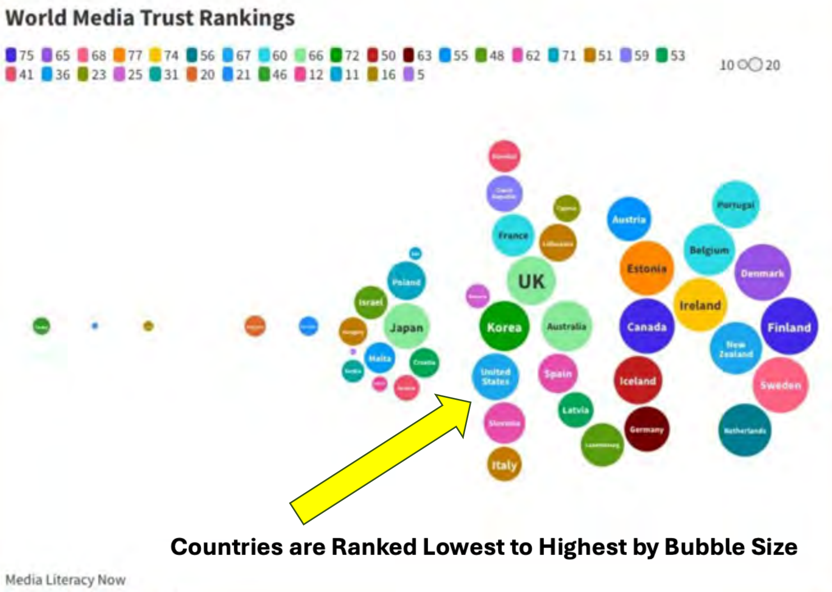
Headline News

Featured

Newsbeat

New Ukrainian Law Limits Military Mobilization To Less Than 500,000, Says Member Of Parliament

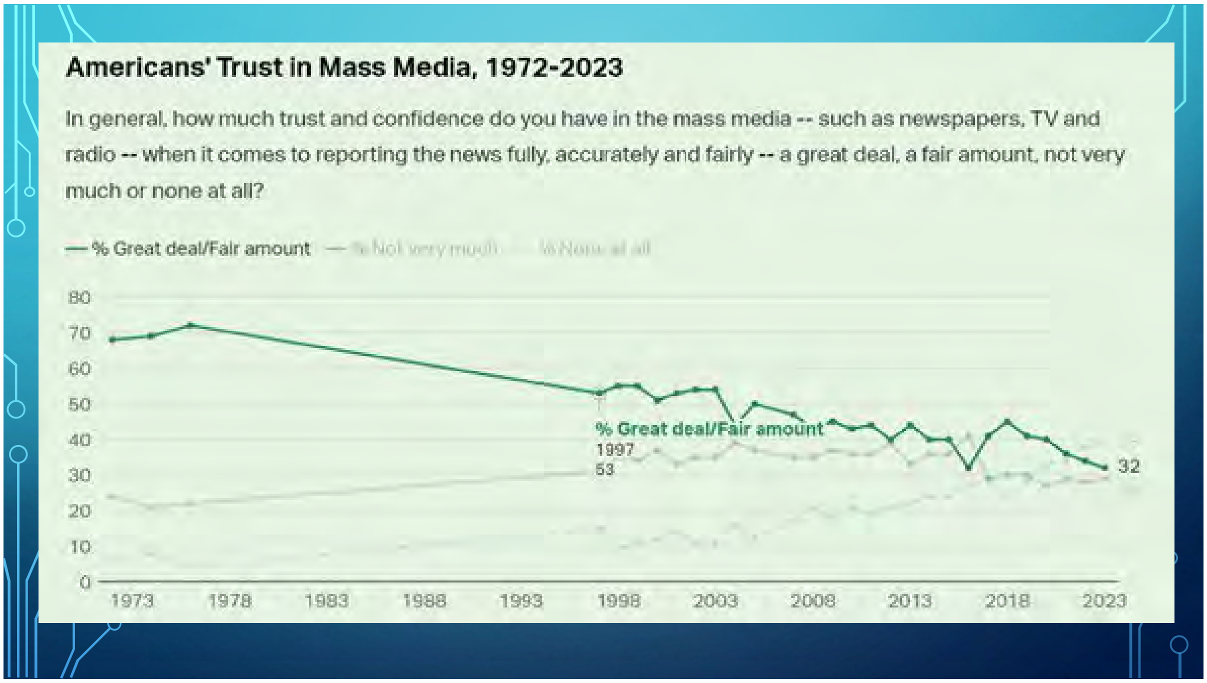
Source: The New York Times



World Media Trust Rankings

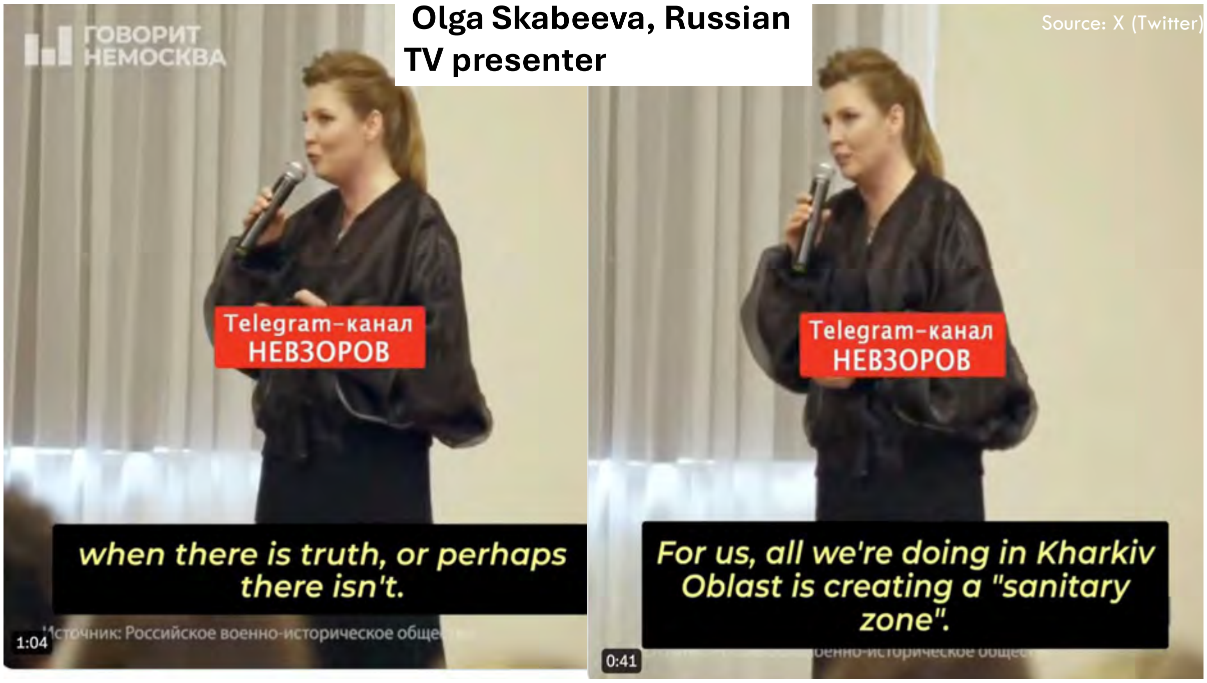
Countries are Ranked Lowest to Highest by Bubble Size

Media Literacy Now



Americans' Trust in Mass Media, 1972-2023

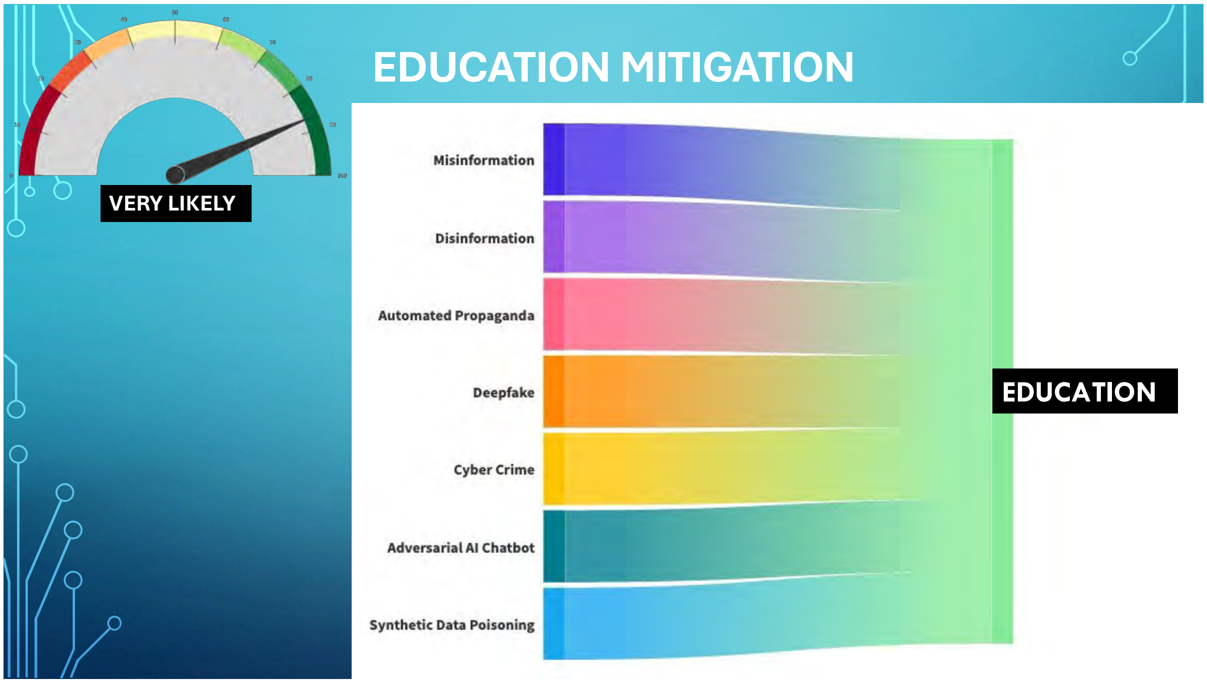
In general, how much trust and confidence do you have in the mass media -- such as newspapers, TV and radio -- when it comes to reporting the news fully, accurately and fairly -- a great deal, a fair amount, not very much or none at all?



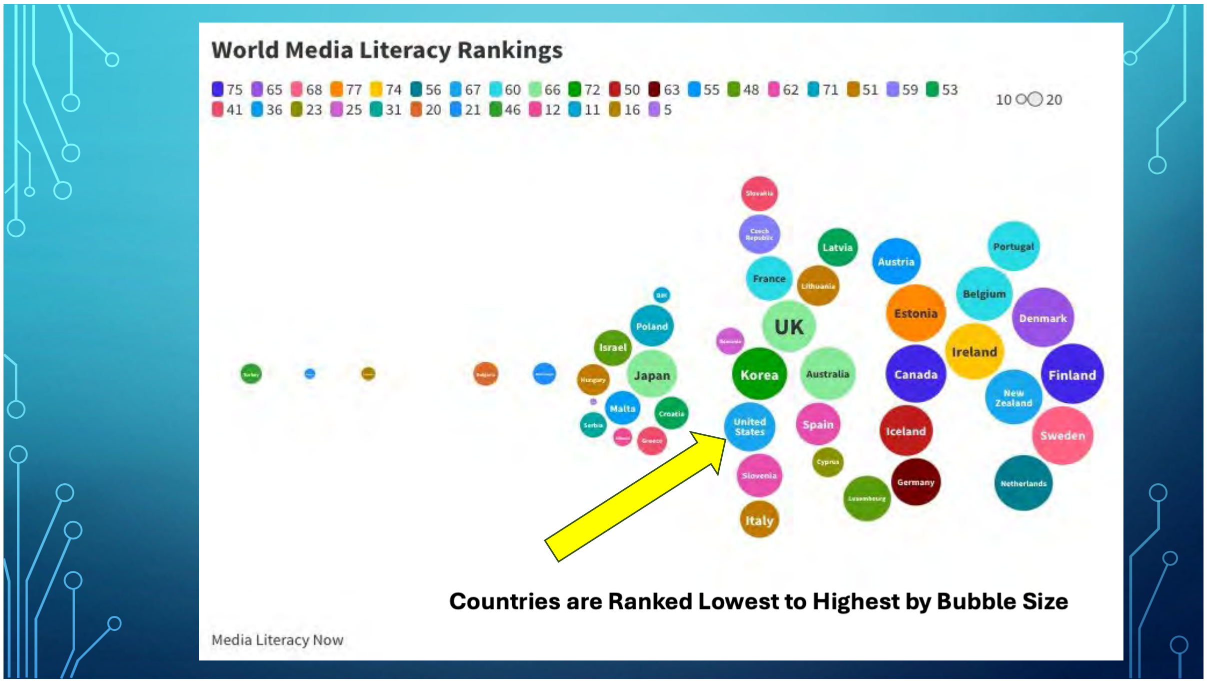
Olga Skabeeva, Russian TV presenter

Source: X (Twitter)

ГОВОРИТ НЕМОСКВА
Telegram-канал НЕВЗОРОВ
when there is truth, or perhaps there isn't.
1:04
Источник: Российское военно-историческое обще
Telegram-канал НЕВЗОРОВ
For us, all we're doing in Kharkiv Oblast is creating a "sanitary zone".
0:41
оенно-историческое общес



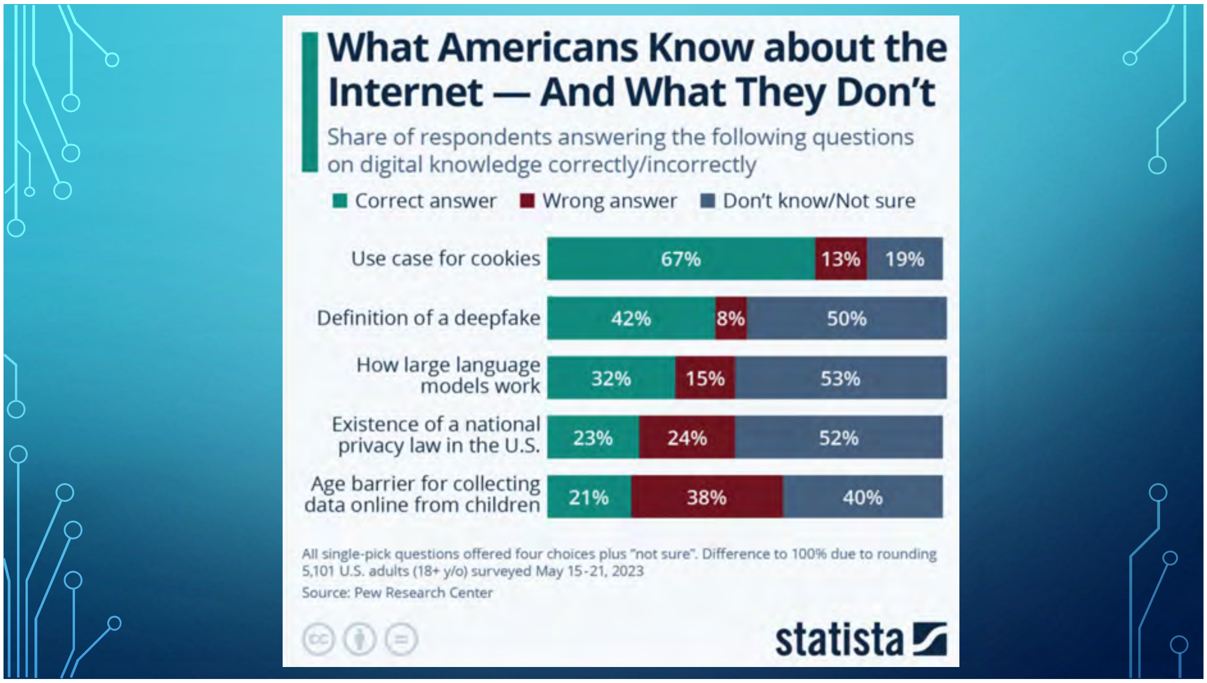
EDUCATION MITIGATION



World Media Literacy Rankings

Countries are Ranked Lowest to Highest by Bubble Size

Media Literacy Now

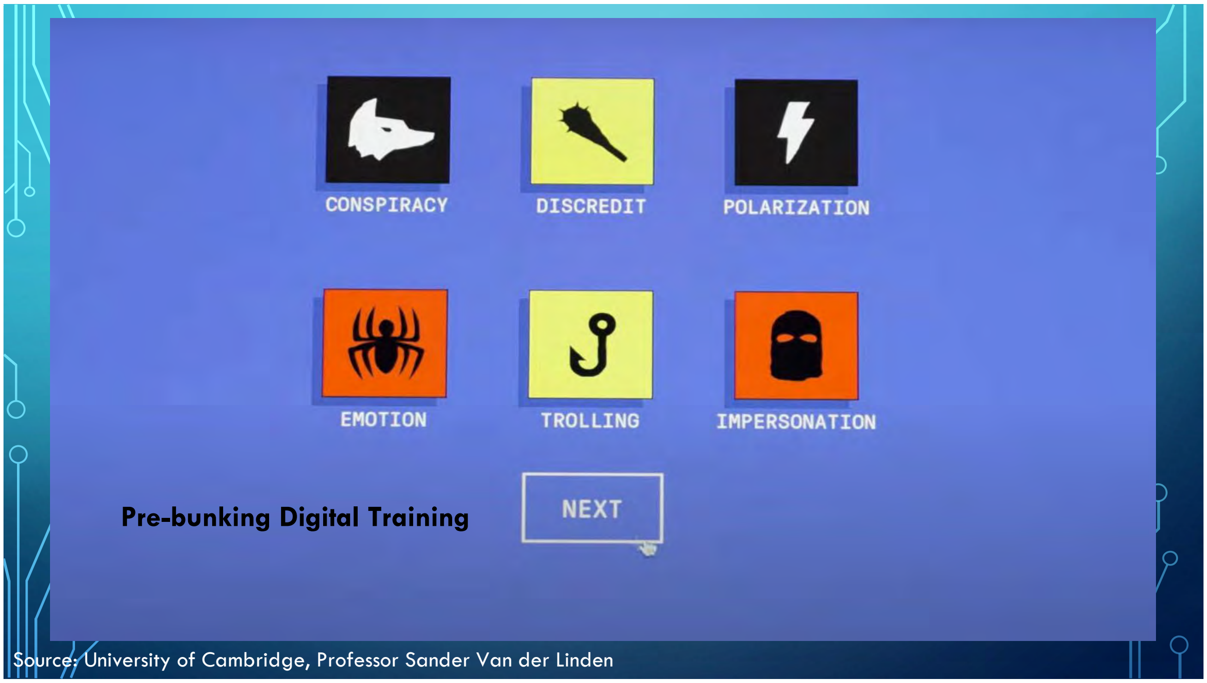


What Americans Know about the Internet — And What They Don't

Share of respondents answering the following questions on digital knowledge correctly/incorrectly

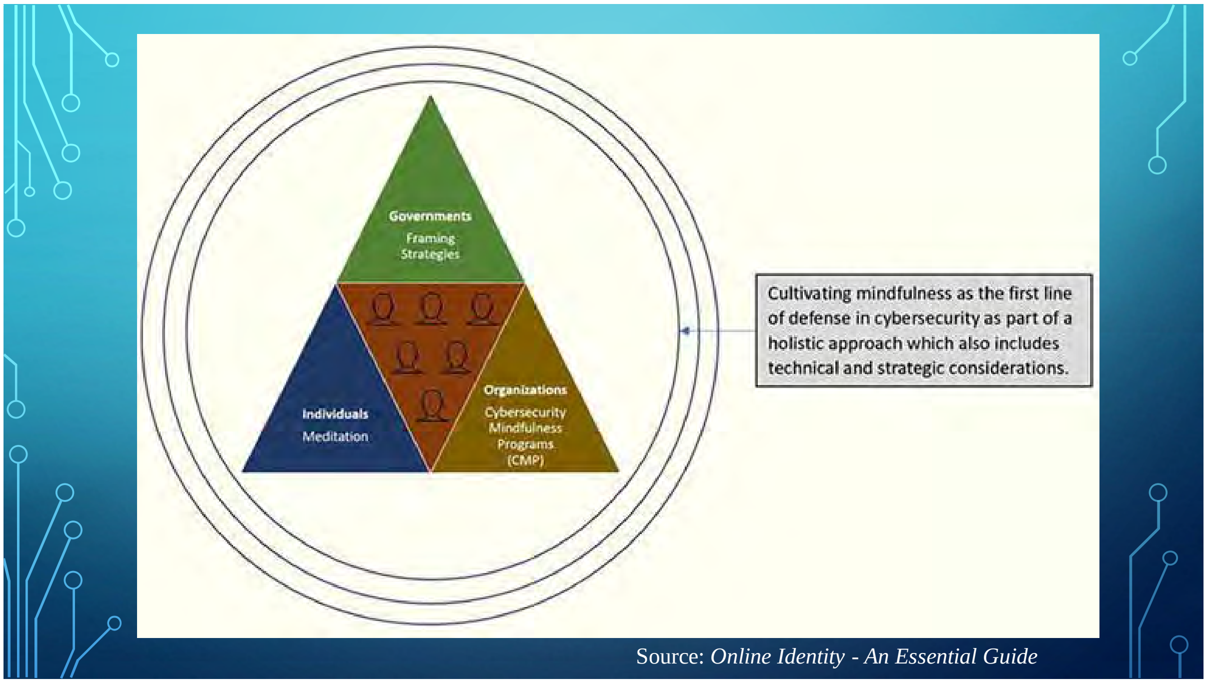
All single-pick questions offered four choices plus "not sure". Difference to 100% due to rounding 5,101 U.S. adults (18+ y/o) surveyed May 15-21, 2023

Source: Pew Research Center



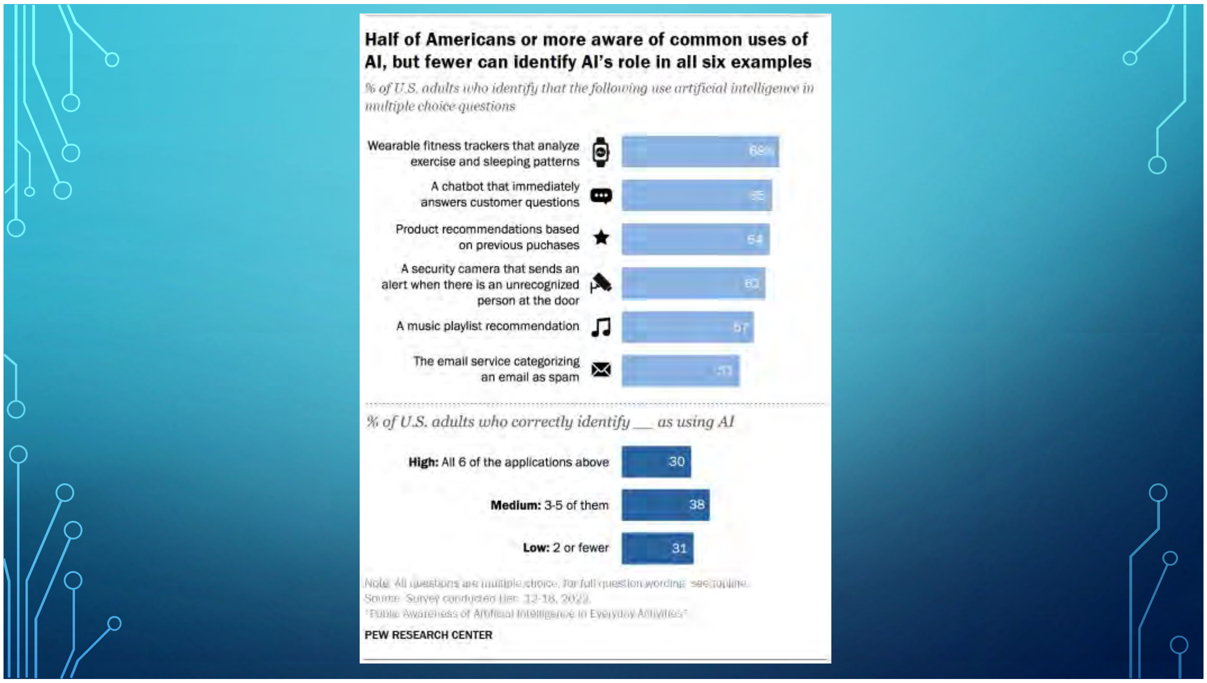
Pre-bunking Digital Training

Source: University of Cambridge, Professor Sander Van der Linden



Cultivating mindfulness as the first line of defense in cybersecurity as part of a holistic approach which also includes technical and strategic considerations.

Source: *Online Identity - An Essential Guide*



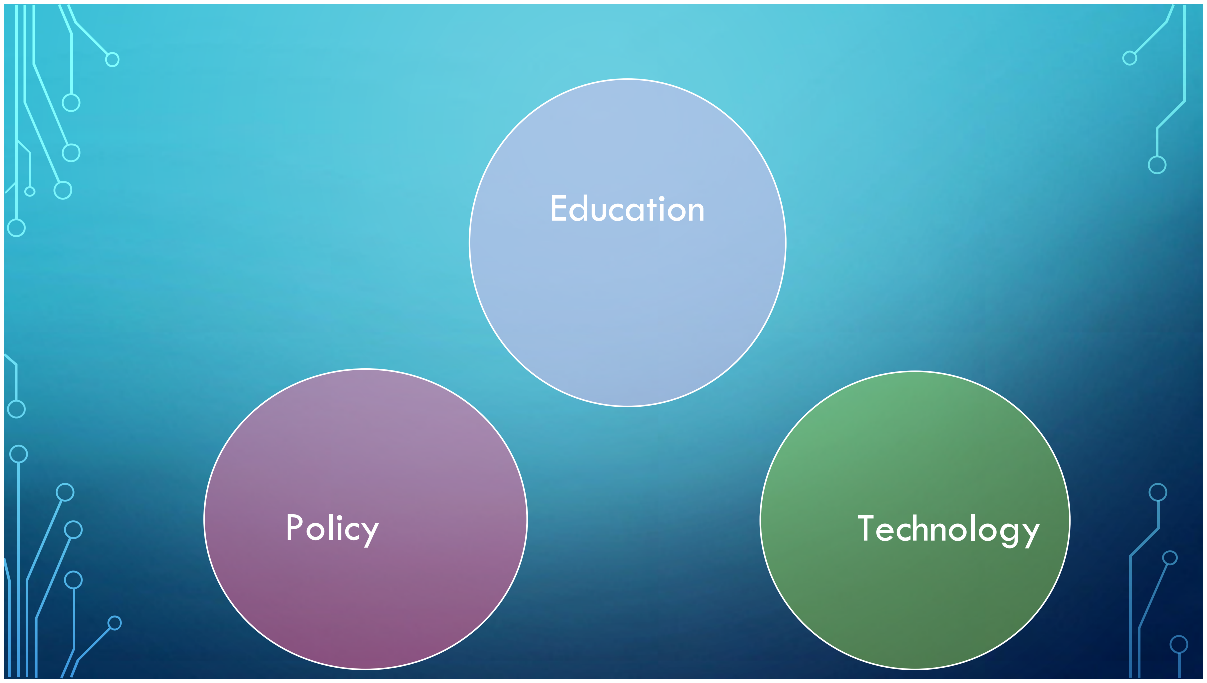
Half of Americans or more aware of common uses of AI, but fewer can identify AI's role in all six examples

% of U.S. adults who identify that the following use artificial intelligence in multiple choice questions

% of U.S. adults who correctly identify __ as using AI

Note: All questions are multiple choice. For full question wording, see topline.
Source: Survey conducted Dec. 12-18, 2022.
"Public Awareness of Artificial Intelligence in Everyday Activities"

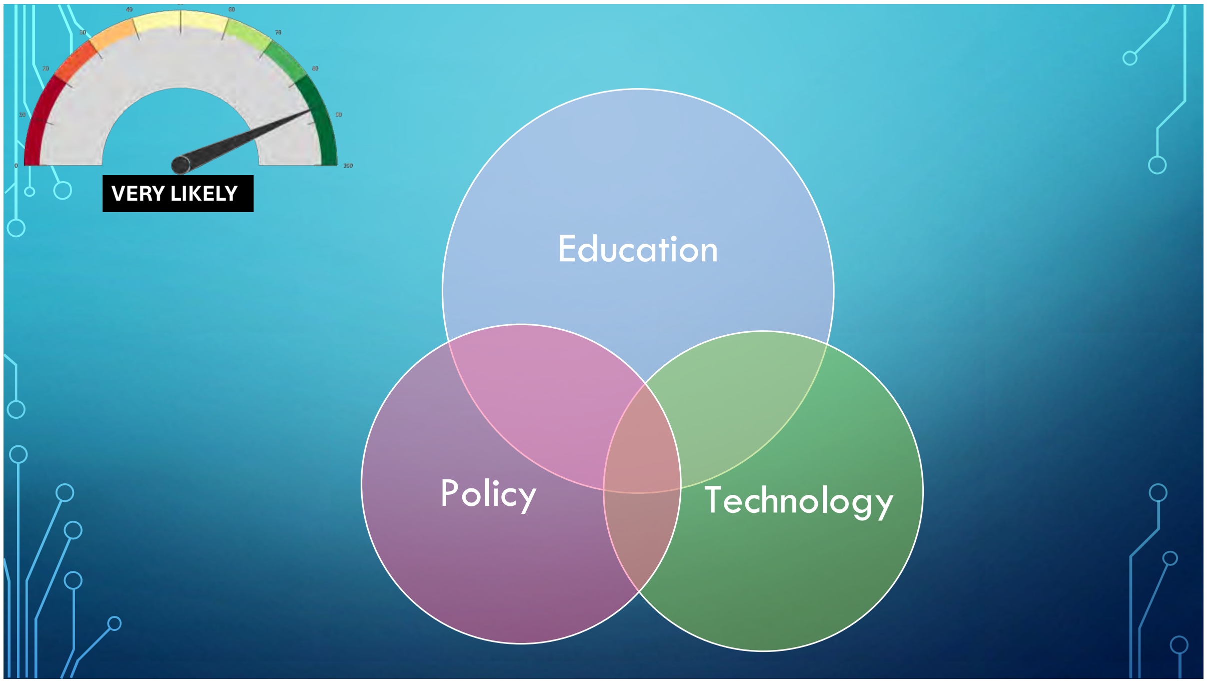
PEW RESEARCH CENTER



Education

Policy

Technology

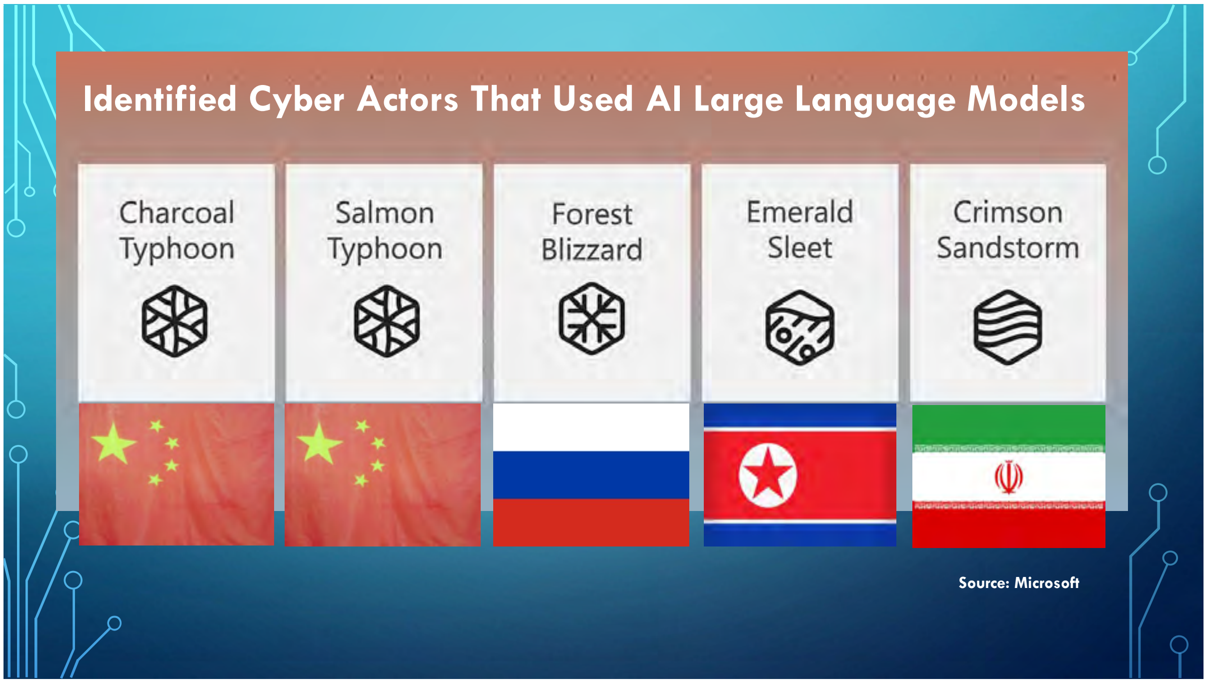


VERY LIKELY

Education

Policy

Technology



Identified Cyber Actors That Used AI Large Language Models

Charcoal Typhoon

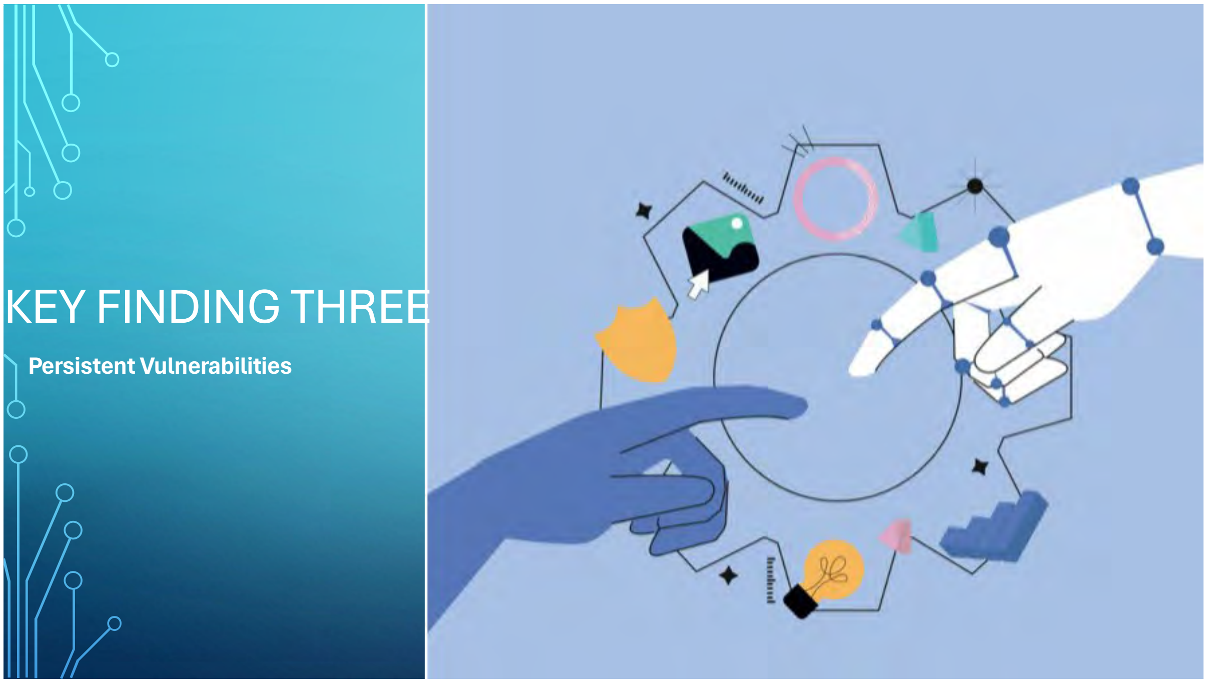
Salmon Typhoon

Forest Blizzard

Emerald Sleet

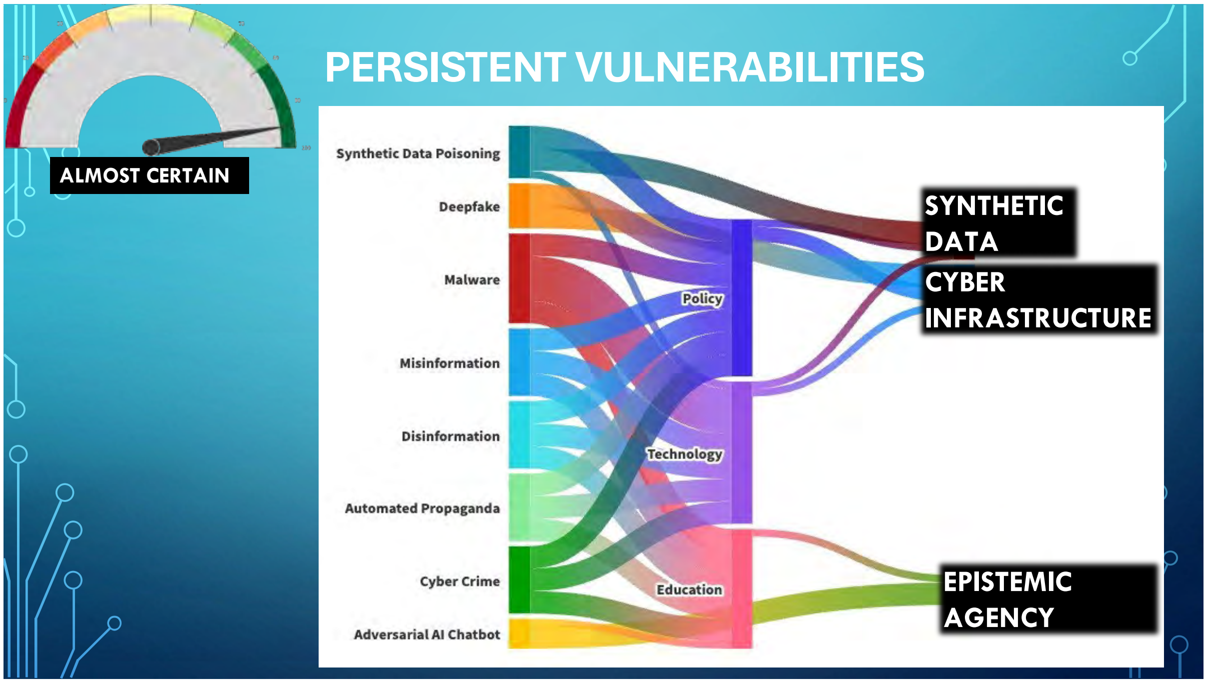
Crimson Sandstorm

Source: Microsoft

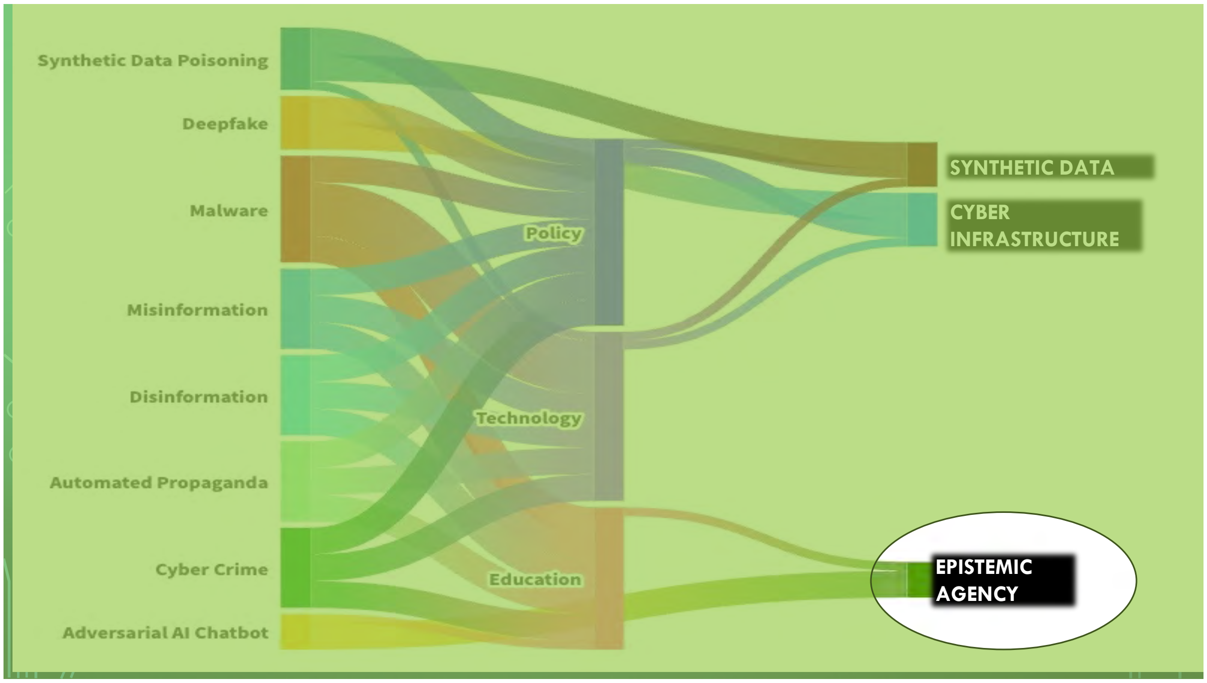


KEY FINDING THREE

Persistent Vulnerabilities

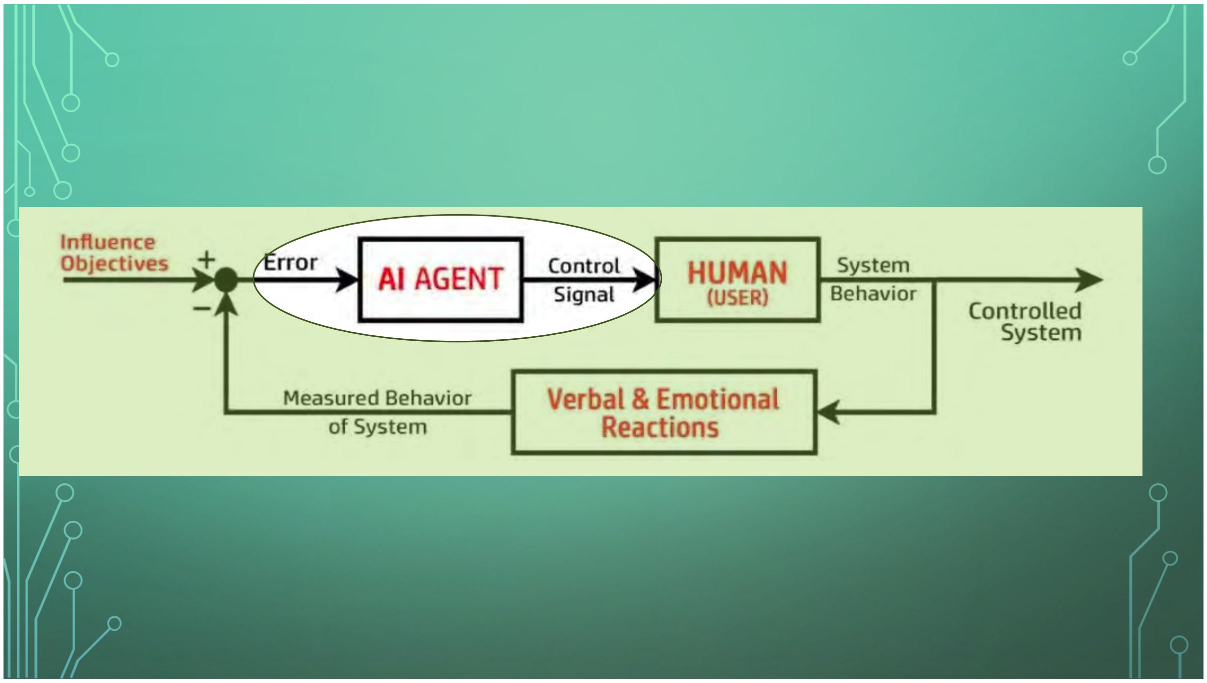


PERSISTENT VULNERABILITIES



ALMOST CERTAIN

Source: Dall-E



Influence Objectives

+

−

Error

AI AGENT

Control Signal

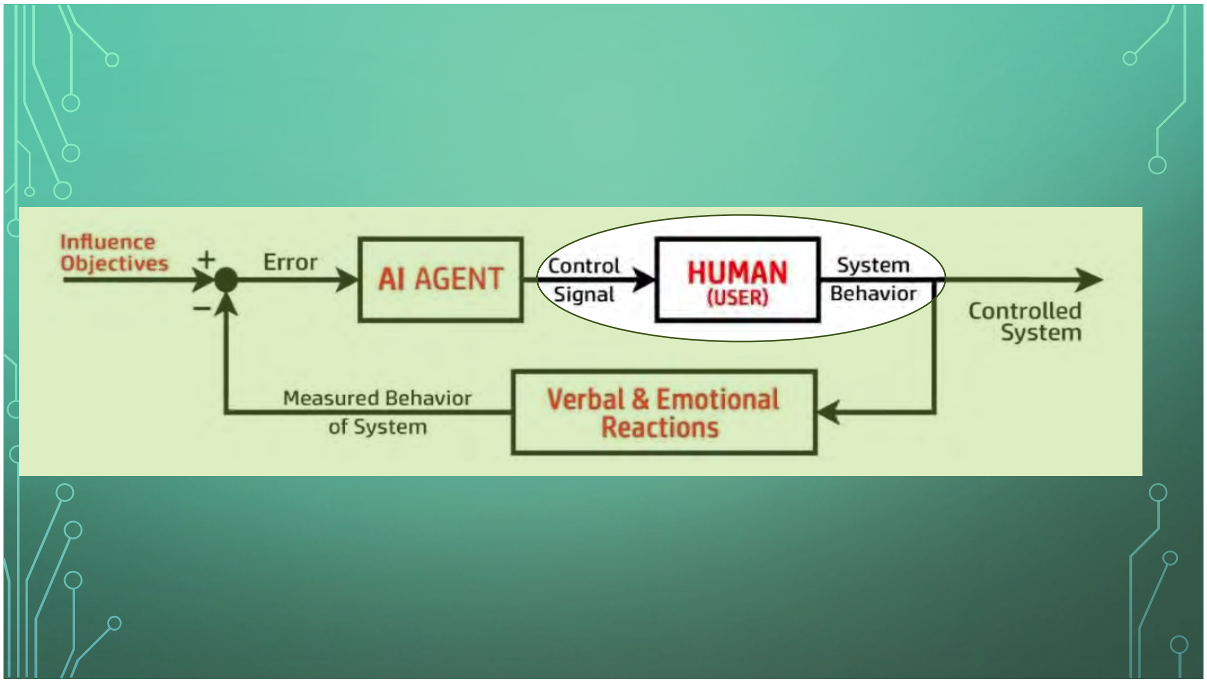
HUMAN (USER)

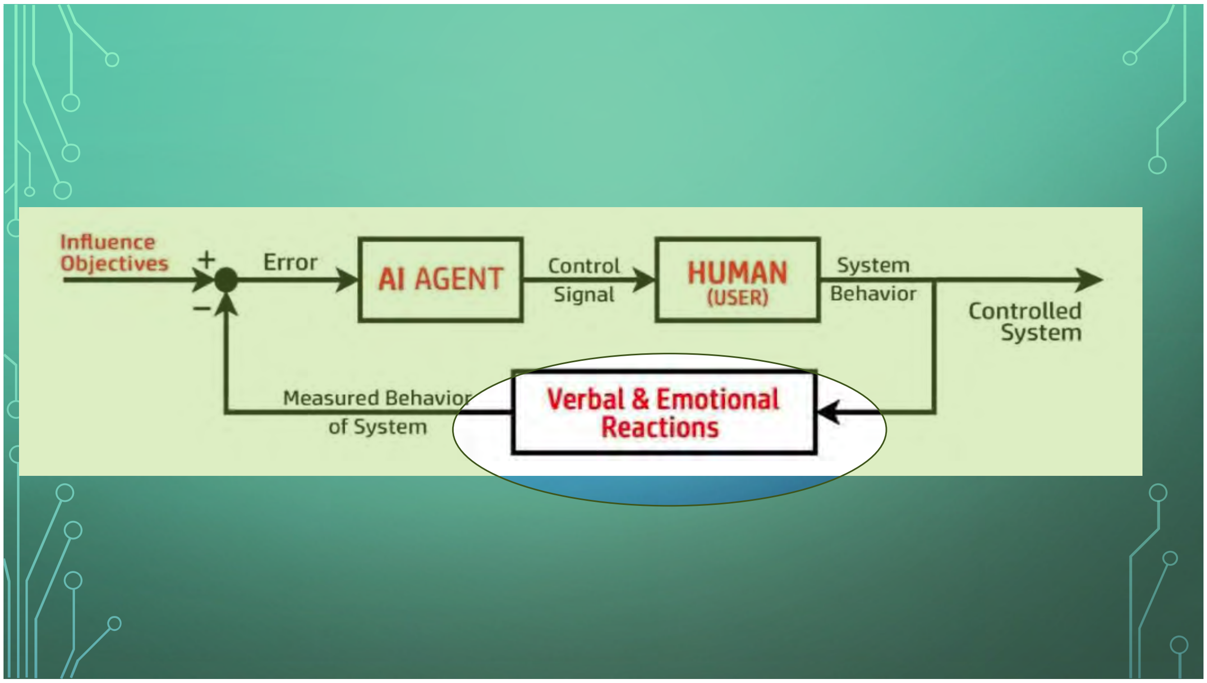
System Behavior

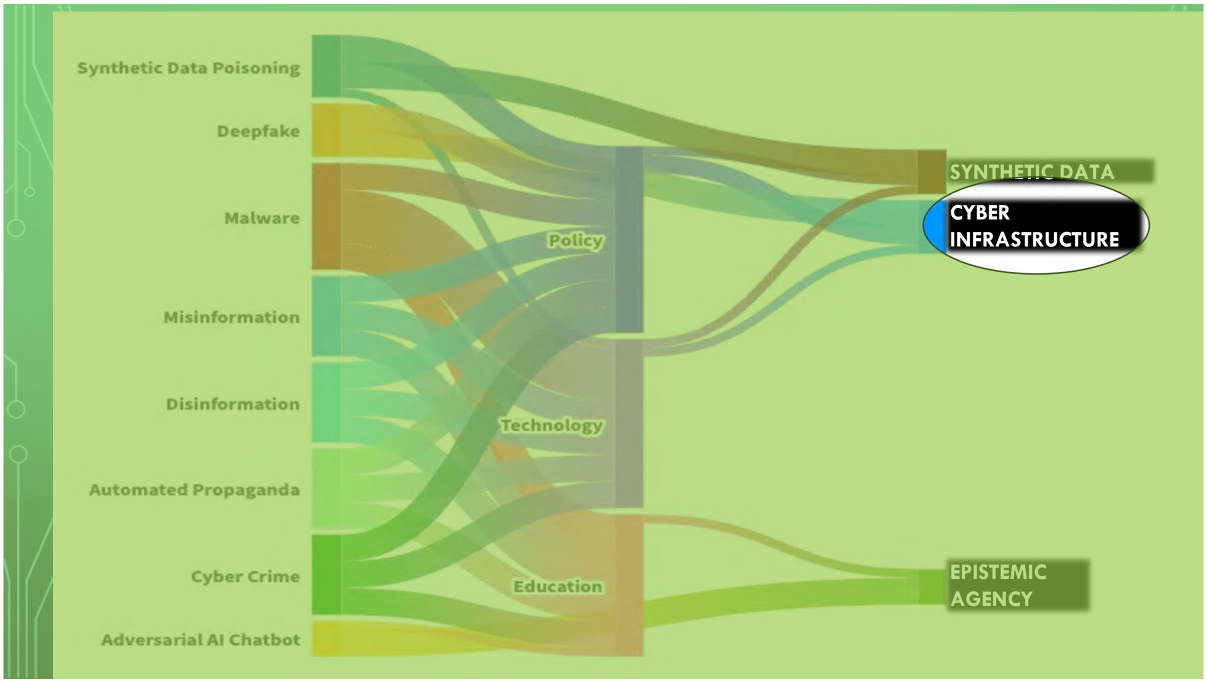
Controlled System

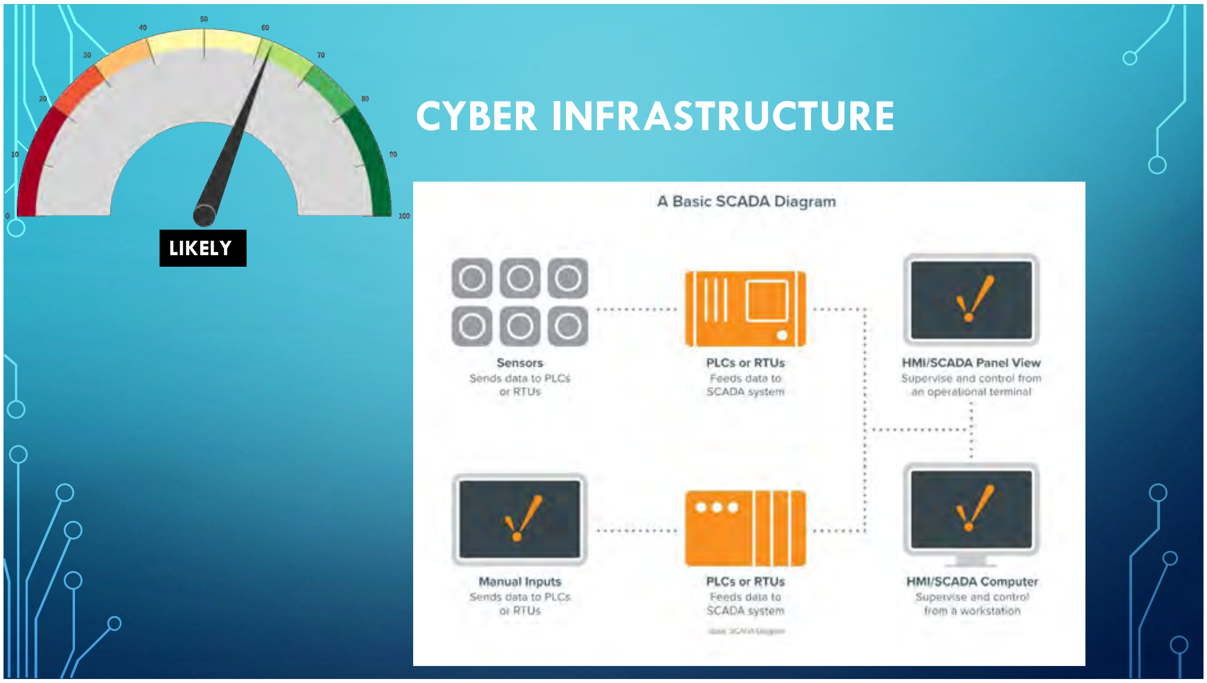
Measured Behavior of System

Verbal & Emotional Reactions


Influence Objectives
+
−
Error
AI AGENT
Control Signal
HUMAN (USER)
System Behavior
Controlled System
Measured Behavior of System
Verbal & Emotional Reactions


Influence Objectives
+
−
Error
AI AGENT
Control Signal
HUMAN (USER)
System Behavior
Controlled System
Measured Behavior of System
Verbal & Emotional Reactions





CYBER INFRASTRUCTURE

A Basic SCADA Diagram

Sensors
Sends data to PLCs or RTUs

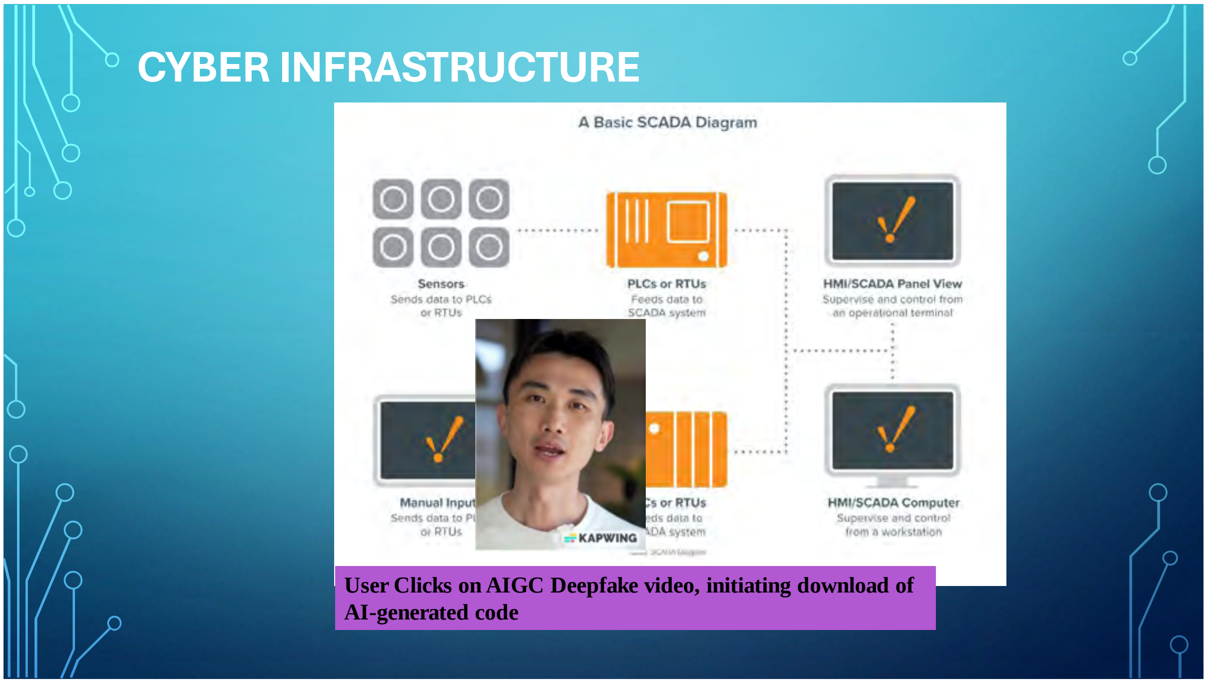
PLCs or RTUs
Feeds data to SCADA system

HMI/SCADA Panel View
Supervise and control from an operational terminal

Manual Inputs
Sends data to PLCs or RTUs

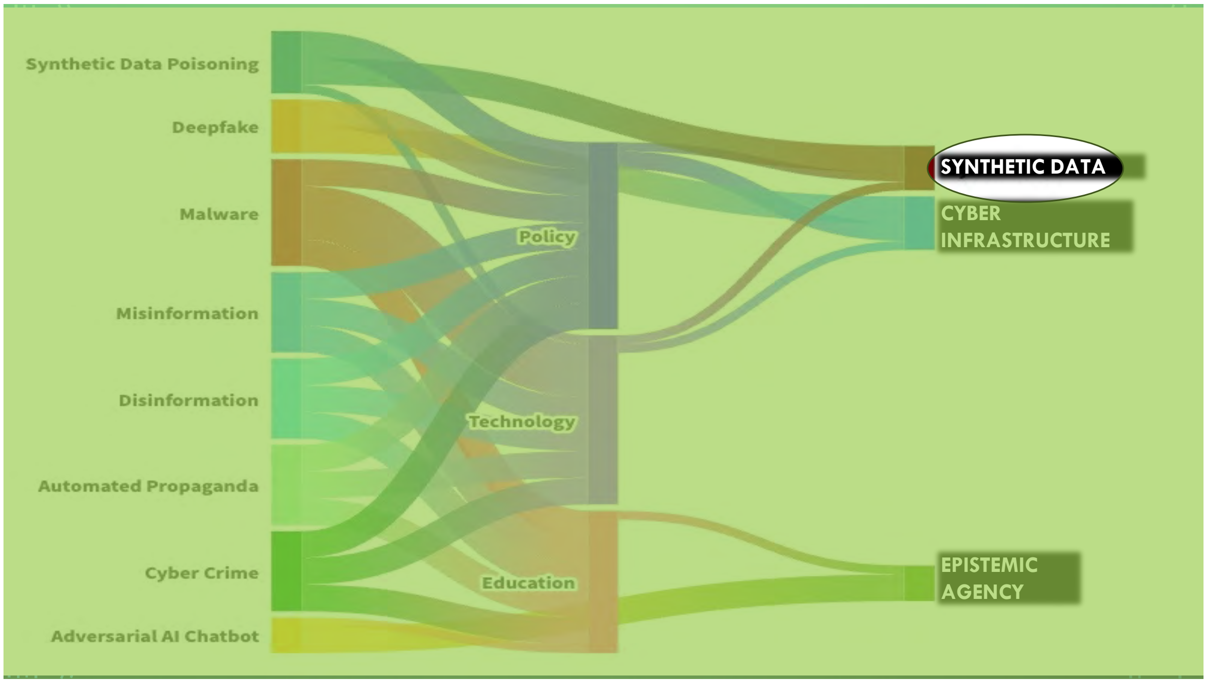
PLCs or RTUs
Feeds data to SCADA system

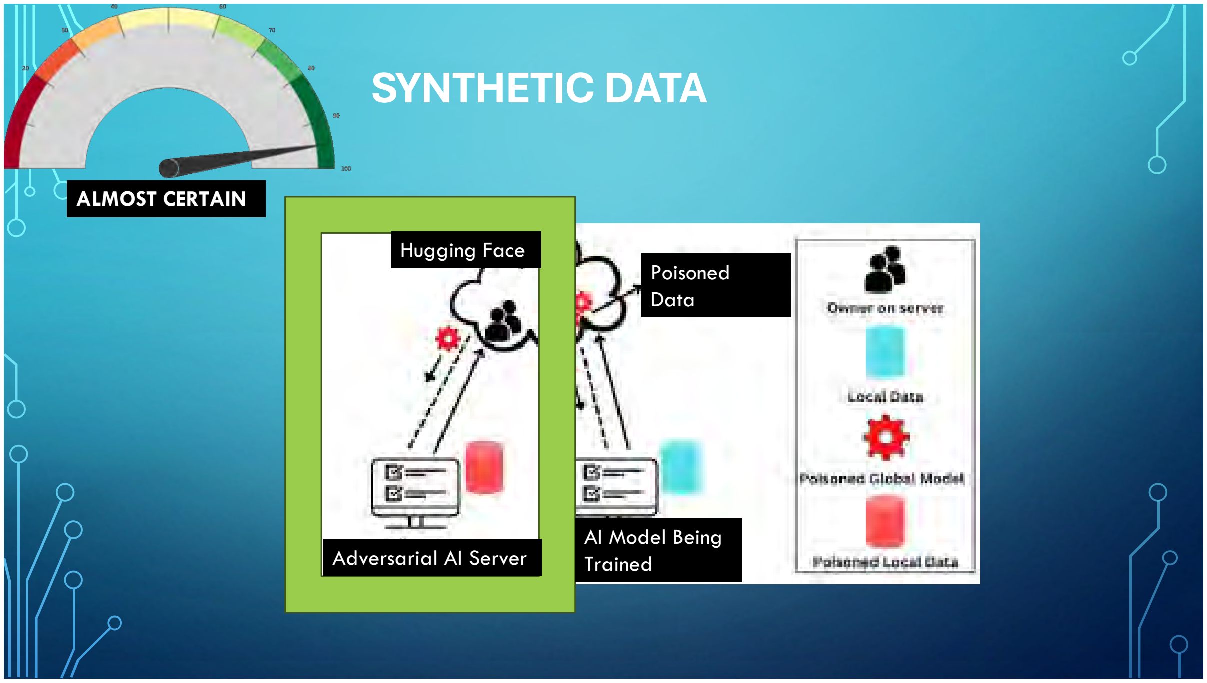
HMI/SCADA Computer
Supervise and control from a workstation



CYBER INFRASTRUCTURE

User Clicks on AIGC Deepfake video, initiating download of AI-generated code





SYNTHETIC DATA

ALMOST CERTAIN

Hugging Face

Poisoned Data

Adversarial AI Server

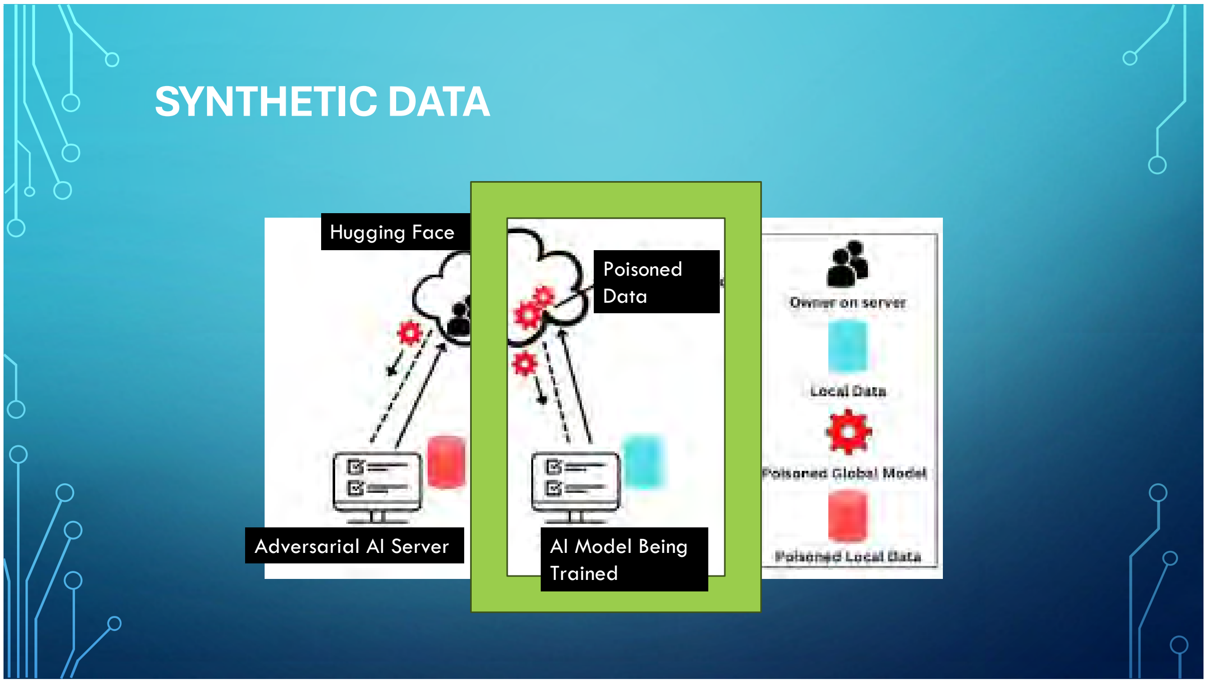
AI Model Being Trained

Owner on server

Local Data

Poisoned Global Model

Poisoned Local Data



SYNTHETIC DATA

Hugging Face

Poisoned Data

Adversarial AI Server

AI Model Being Trained

Owner on server

Local Data

Poisoned Global Model

Poisoned Local Data

Hugging Face

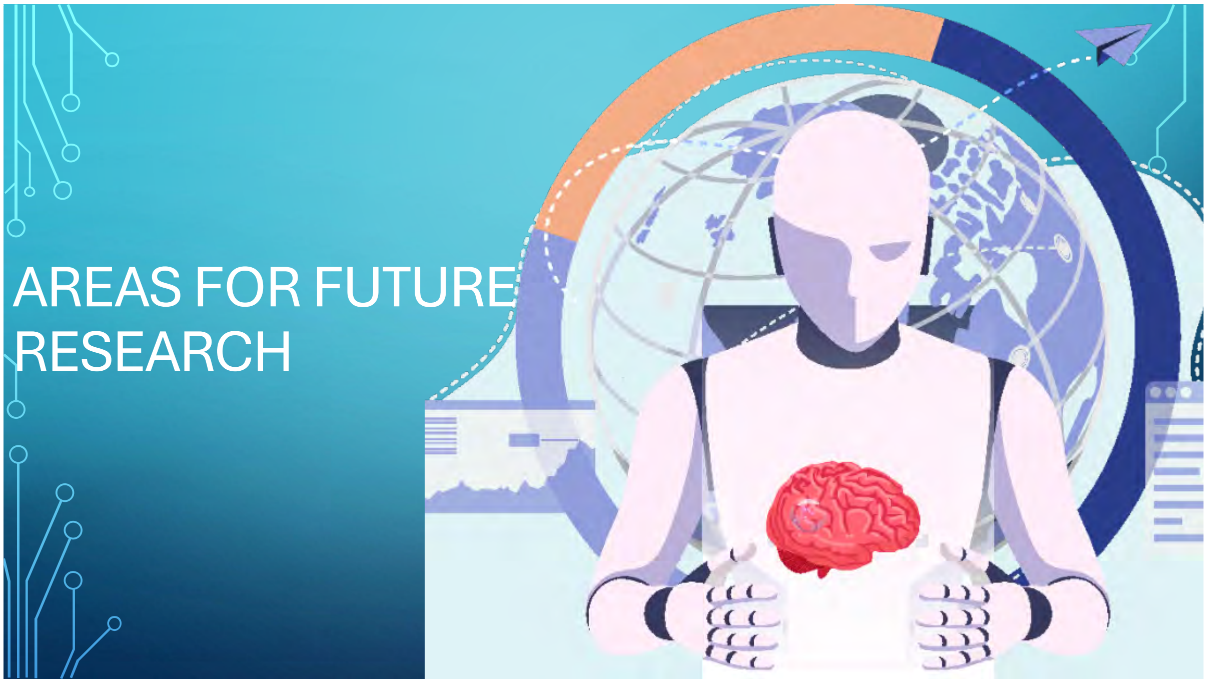
Poisoned Data

Adversarial AI Server

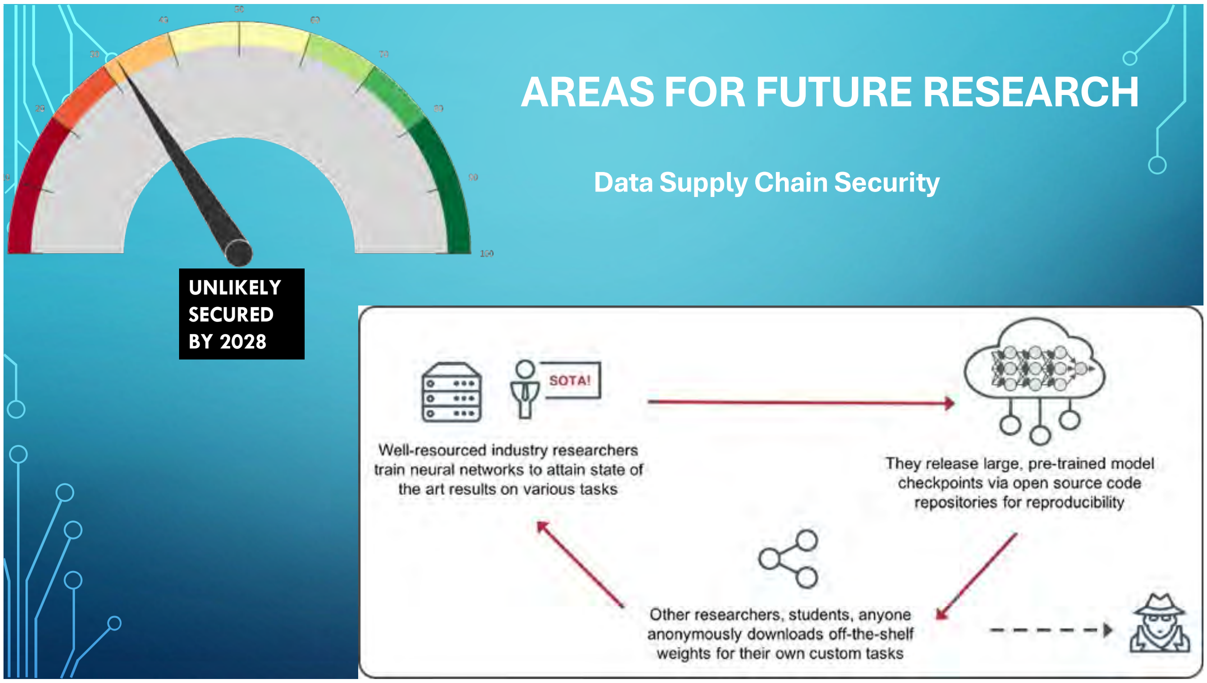
AI Model Being Trained

U.S. Air Force Conducts World's First Ever Human VS AI Dogfight

Source Defense News



AREAS FOR FUTURE RESEARCH



AREAS FOR FUTURE RESEARCH

Data Supply Chain Security

UNLIKELY SECURED BY 2028

Well-resourced industry researchers train neural networks to attain state of the art results on various tasks

They release large, pre-trained model checkpoints via open source code repositories for reproducibility

Other researchers, students, anyone anonymously downloads off-the-shelf weights for their own custom tasks

AREAS FOR FUTURE RESEARCH

ALMOST CERTAIN to USE AIGC by 2028

Nations

Non-Aligned

Illustration: James Marshal, Source: Wired

GAO

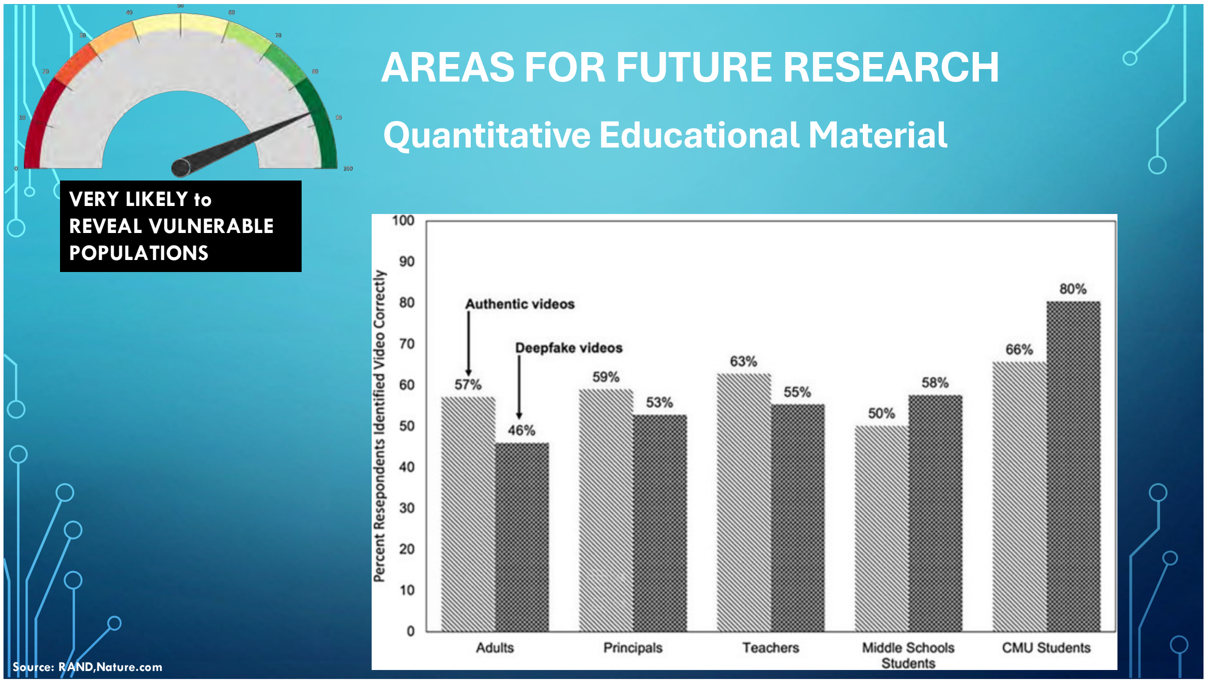
United States Government Accountability Office

Report to Congressional Requesters

September 2023

VIOLENT EXTREMISM

Agencies' and Financial Institutions' Efforts to Link Financing to Domestic Threats



AREAS FOR FUTURE RESEARCH

Quantitative Educational Material

VERY LIKELY to REVEAL VULNERABLE POPULATIONS

Source: RAND,Nature.com

RAMIREZ
LAS VEGAS REVIEW-JOURNAL for THE WASHINGTON POST 2024©
This ADVANCED ARTIFICIAL INTELLIGENCE WILL EXPOSE A.I. WITH DECEITFUL or MALICIOUS INTENTIONS.
HOW DO WE KNOW IT'S NOT LYING?

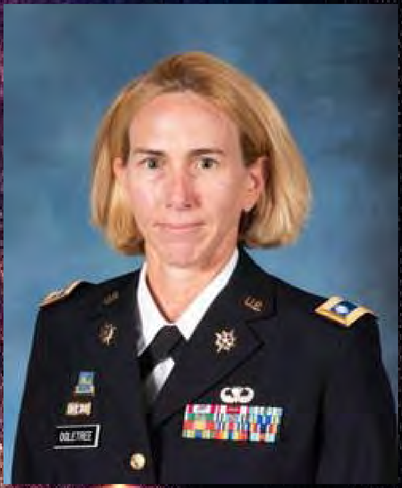
DISCUSSION & QUESTIONS

Source: Washington Post

Mr. Tom Jackson

COL “Marc” Richardson

LTC Charlie Moss



LTC Kate Ogletree

CDR Robert Liberato

ERGO SUM
MACHINA