Federal Register Notice 87 FR 15274, https://www.federalregister.gov/documents/2022/03/17/2022-05683/request-for-information-on-federal-priorities-for-information-integrity-research-and-development, May 15, 2022

Request for Information on Federal Priorities for Information Integrity Research and Development

Special Competitive Studies Project (SCSP)

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

SPECIAL COMPETITIVE STUDIES PROJECT

Subject: RFI Response: Information Integrity R&D

From: Special Competitive Studies Project (SCSP) Staff

To: Networking and Information Technology Research and Development (NITRD), National Coordination Office (NCO), and National Science Foundation (NSF)

Date:

Introduction and Background on SCSP

The Special Competitive Studies Project (SCSP) is a non-profit organization committed to strengthening America's long term competitiveness for a future where artificial intelligence (AI) and other emerging technologies reshape our national security, economy, and society.

SCSP's work concentrates on six policy areas where AI and emerging technologies play a critical role: foreign policy, intelligence, defense, economy, society, and future technology platforms. The question of information integrity cuts across all these subject areas.

Using the expertise of our staff and insights gained from SCSP's engagements with various stakeholders, we offer suggestions responding to question two of the request for information: *preserving information integrity and mitigating the effects of information manipulation*.

The challenges surrounding the information ecosystem, particularly information integrity and manipulation, are complex issues that must be addressed in a multidisciplinary, holistic manner before further eroding the information space. Our response consequently emphasizes prioritizing baseline understandings of technique effectiveness; ecosystem mapping; developing capabilities to counter information influence and manipulation at scale; overcoming the Liars' Dividend; increased emphasis on human involvement in the information environment; research on differential impact internationally; and increased linguistic and cultural capabilities for managing various information ecosystems. Each is described in greater detail below.

The SCSP staff are ready to engage with NITRD, NCO, and NIST regarding any questions or desire for further discussion.

**Disclaimer – The comments and suggestions provided in this RFI response are those of SCSP team members and do not necessarily represent the views of the SCSP Board, its leadership, or the organization as a whole.*

Information Requested by the RFI

2. *Preserving information integrity and mitigating the effects of information manipulation:* Strategies for protecting information integrity must integrate the best technical, social, behavioral, cultural, and equitable approaches. These strategies should accomplish a range of objectives including to detect information manipulation, discern the influence mechanisms and the targets of the influence activities, mitigate information manipulation, assess how individuals and organizations are likely to respond, and build resiliency against information manipulation.

    a. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives?

Responses to the RFI

1. Common Understanding of the Effectiveness of Existing Tools and Techniques

Recommendation: Research changes and differences in behavioral response rates to best practice information manipulation interventions and quantitatively assess the effects of different methods.

Intended Outcome: Scientifically rigorous evidence of the effectiveness of various techniques for responding to information manipulation that will act as a basis for creating strategies and response frameworks.

A primary gap that the Information Integrity R&D Working Group should focus on is the establishment of a common understanding of the effectiveness of existing tools and techniques. Research in the field primarily investigates the effectiveness of mitigation tools such as warning labels, source alerts, and fact-checking.[1] Such research provides a starting point, but building strong strategies to preserve information integrity and

[1] See e.g., Jack Nassetta & Kimberly Gross, State Media Warning Labels Can Counteract the Effects of Foreign Misinformation, Harvard Kennedy School Misinformation Review (2020); Jason Ross Arnold, et al., Source Alerts Can Reduce the Harms of Foreign Disinformation, Harvard Kennedy School Misinformation Review (2021).

mitigate the effects of information manipulation requires a broader baseline understanding of the countermeasures being implemented.

Recent reports, such as those published by Carnegie's Partnership for Countering Influence Operations, show the field lacks evidence-based research for understanding the impact of both influence operations and countermeasures.[2] Sponsoring research into how user behavior changes as a result of different influence operations and countermeasures to provide a scientifically rigorous assessment of the effectiveness of 'best-practice' tools would validate and ideally strengthen the effects of future strategies.

2. Limited Ecosystem Collaboration

Recommendation: Establish a neutral third party to aggregate data and act as an information broker between platforms.

Intended Outcome: Encourage a formalized incident sharing system for collating information related to information integrity concerns and information manipulation.

The online information ecosystem[3] consists largely of social media platforms. Those platforms are siloed and each company has access to their unique information, limiting the shared understanding of incidents related to information manipulation across platforms (large and small). While it is necessary to keep certain information internal for proprietary business practices, new challenges in the information space require enhanced transparency in a way that promotes a safe ecosystem without undermining businesses. During SCSP staff's engagements with experts in this field, the experts described larger platforms as informally increasing coordination during the Russia-Ukraine War to counter foreign influence operations. To best support a healthy ecosystem that includes both large and small companies, a system for incident sharing, including best practices for sharing without revealing proprietary information and protecting individual privacy, that incorporates as many platforms as possible is recommended.

To best detect, discern, and mitigate information manipulation networks, SCSP recommends establishing a formal centralized system of incident reporting via a neutral third party to cover the information ecosystem. Similar to the function played by the

[2] See e.g., Jon Bateman, et al., Measuring the Efficacy of Influence Operations Countermeasures: Key Findings and Gaps From Empirical Research, Carnegie Endowment for International Peace (2021); Jon Bateman, et al., Measuring the Effects of Influence Operations: Key Findings and Gaps From Empirical Research, Carnegie Endowment for International Peace (2021).

[3] SCSP recommends that the Information Integrity R&D Working Group define the information ecosystem as it relates to these priorities, given the contributions of other factors including digital financial services and the Internet of Things to the information ecosystem.

MITRE Corporation in aggregating incident data for the aviation industry,[4] a neutral third party would act as an information broker and data manager for data cleared in relation to identified and reported incidents. The establishment of such a system would encourage a set of norms for platforms to report, share, and act when incidents are spotted in addition to providing a centralized point of contact for handling active cases of information manipulation.

3. Capabilities for Fighting Autonomous Disinformation at Speed and Scale

Recommendation: Develop a national framework for how to combine technical tools and human capabilities to combat autonomous disinformation at speed and scale.

Intended Outcome: A strong, multi-method approach for preempting and protecting against the newest wave of information manipulation and influence – autonomous disinformation.

AI-enabled autonomous disinformation, or information manipulation, is here. How to build resiliency and defend against the barrage of information at the same speed and scale as it arrives – or even faster – remains unclear. The National Security Commission on Artificial Intelligence (NSCAI)'s Final Report warned that the "U.S. government is not prepared to defend the United States in the coming artificial intelligence (AI) era."[5] The report also noted that "AI and associated technologies will increase the magnitude, precision, and persistence of adversarial information operations" through AI-generated messages that are nearly indistinguishable from authentic messages, the targeting of specific audiences, and the proliferation of malign information through platforms.[6]

A cohesive national framework combining capacities across public and private sectors, while also seeking the input of subject matter experts, would facilitate countering autonomous disinformation in the future. The process of creating a national strategy should identify the strengths and weaknesses of current tools and techniques, to best identify priority areas for strengthening such a strategy.

Based on expert consultations, SCSP suggests a national framework explore the necessity of psychological, technical, social, and cultural components. Most importantly, this holistic framework would equally incorporate human and machine capabilities in order to maximize the reach of existing tools and techniques. With the human-machine team as a base, the strategic framework would combine best practice tools and fact-checking with new techniques being developed to build resilience.[7]

---

[4] Marlis McCollum, MITRE Adds A Special Element Of Trust To Data Sharing And Analysis, MITRE (2019).
[5] National Security Commission on Artificial Intelligence, Final Report at 45 (2021).
[6] Id. at 47-48.
[7] One example of a tool developed to help educate the public on media literacy and misinformation is Bad News. See https://www.getbadnews.com/#intro.

4. Overcoming the Liars' Dividend

Recommendation: Sponsor research into labeling mechanisms and public safety warnings about information manipulation, particularly those that prevent a decrease in trust in credible information.

Intended Outcome: Provide an evidence-based, trust-building suite of techniques for mitigating information manipulation to government, academia and the private sector.

The Liars' Dividend increasingly exacerbates diminishing trust in information. Originally defined by Chesney and Citron, the Liars' Dividend refers to the situation in which the benefit to malicious actors "flows, perversely, in proportion to success in educating the public about the dangers of deep fakes."[8] In other words, the more individuals are educated about the dangers of deepfakes, the more likely they are to doubt the information they come across regardless of the truth of the information. Consequently, some actors gain traction for misleading or manipulative narratives by denying the credibility of online content.

Research into behavioral responses to deepfake warning labels is limited. A recent study shows that the majority of individuals are unable to differentiate between unaltered and deepfake videos when given a warning that content may have been altered.[9] The preliminary study noted that improved deepfake detection, combined with the human inability to differentiate when given media warnings, may further diminish trust in the media[10] – contributing to the Liars' Dividend.

Chapter 1 of the NSCAI Final Report recommends that The White House Office of Science and Technology (OSTP) take the lead on studying AI and complementary technologies for certifying content authenticity and provenance.[11] Building upon NSCAI's recommendation, SCSP suggests the Information Integrity R&D Working Group at NITRD sponsor research regarding behavioral responses to user interface designs warning about synthetic media and techniques that also build trust in content moderators issuing labels. Identifying ways to provide credible information and warnings about disinformation without further deteriorating societal trust in content is critical to building a strong information ecosystem. The research would also support the proposed task force by providing insights about additional complementary tools for how labeling mechanisms and public safety warnings pertaining to information manipulation impact user behavior. Such research should pool expertise from government, academia, and the private sector to maximize insight.

---

[8] Danielle K. Citron & Robert Chesney, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review at 1785 (2019).
[9] Andrew Lewis, et al, Do Content Warnings Help People Spot a Deepfake? Evidence from Two Experiments, The Royal Society (2022).
[10] Id. at 17.
[11] National Security Commission on Artificial Intelligence, Final Report at 49 (2021).

5. Incorporating the Human Element

Recommendation: Send out a call for papers analyzing the role of human influence methods and information manipulation.

Intended Outcome: Develop a complementary understanding of how human operations interact with digital ecosystems to impact information integrity and resilience.

Research into information campaigns continues to focus on the digital domain and social media aspects of influence operations. For example, the Stanford Internet Observatory,[12] The Citizen Lab,[13] Atlantic Council's Digital Forensic Research Lab,[14] Australian Strategic Policy Institute,[15]etc. have focused most of their work on online forms of influence. This platform-centric approach too often overlooks ways in which malign actors manipulate and exploit routine social behavior in the physical world.

One example of the role of humans in influence campaigns is the Australian 2017 Bennelong By-election. The Chinese Communist Party, by using a series of groups and exploiting new digital tools, aimed to exert influence over the votes to target the Turnbull government.[16] Common understanding of how such traditional human influence operations continue to play a role in shaping the information ecosystem is needed for a holistic understanding of the field. This understanding is a requirement for identifying and exploring how traditional actors and methods are enabled by existing and emerging platforms and technologies.

SCSP recommends the Information Integrity R&D Working Group send out a call for papers investigating foreign human influence operations. Topics most beneficial to the field include the tools and techniques that combine human influence operations with information manipulation on platforms, the evolving role of human influence operations, and comparative trends among those targeted for influence activities. The resulting research would strengthen strategies for countering information manipulation, while also broadening common understanding in the field around the role of humans in maintaining – or manipulating – the information ecosystem.

---

[12] Shelby Grossman, et al., Full-Spectrum Pro-Kremlin Online Propaganda about Ukraine: Narratives from Overt Propaganda, Unattributed Telegram Channels, and Inauthentic Social Media Accounts, Stanford Internet Observatory (2022).
[13] John Scott-Railton, et al., CatalanGate: Extensive Mercenary Spyware Operation against Catalans Using Pegasus and Candiru, The Citizen Lab (2022).
[14] Jean-Baptiste Jeangène Vilmer, Information Defense: Policy Measures Taken Against Foreign Information Manipulation, Atlantic Council Digital Forensic Research Lab at 25-27 (2021).
[15] Jacob Wallis & Albert Zhang, Understanding Global Disinformation and Information Operations: Insights from ASPI's new analytic website, Australian Strategic Policy Institute (2022) (Using data sets from Twitter's Information Operations archive).
[16] Alex Joske, Bennelong Byelection: The Influential Network Targeting the Turnbull Government in Bennelong, Sydney Morning Herald (2017).

Information Requested by the RFI

2. *Preserving information integrity and mitigating the effects of information manipulation:* Strategies for protecting information integrity must integrate the best technical, social, behavioral, cultural, and equitable approaches. These strategies should accomplish a range of objectives including to detect information manipulation, discern the influence mechanisms and the targets of the influence activities, mitigate information manipulation, assess how individuals and organizations are likely to respond, and build resiliency against information manipulation.

   b. What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?

Responses to the RFI

1. Limited Research and Visibility of Differential Impact in Other Countries

Recommendation: Prioritize research about the relationship between national information ecosystems (their vulnerabilities and points of fracture) and how malign actors abroad choose strategic targets for exploitation and then propagate those exploits to other nations. This research can help shape domestic policies to make national information ecosystems more resilient.

Intended Outcome: Increased comprehensive research related to the global information ecosystem by gaining greater understanding of national information ecosystems and how targeted attacks on another nation may be an indicator of what to expect within the U.S.

Researchers within the United States and policymakers tend to be domestically focused. In doing so, they may overlook strategic implications of groups targeted in other countries. However, the information integrity in other countries is of strategic value to the United States given that information manipulation and influence propagate across borders.

For example, malign actors increasingly use Latin American and African countries to expand the range of influence operations,[17] many times as test beds before they are launched within the United States. Russia Today recently targeted Latin American

[17] The Global Engagement Center: Leading the United States Government's Fight Against Global Disinformation Threat, Hearing Before the Subcommittee on State Department and USAID Management, International Operations, and Bilateral International Development, of the Senate Committee on Foreign Relations (2020).

countries through the Russian television channel, RT en Español, with disinformation about the Russian Invasion in Ukraine.[18] Researchers at the Atlantic Council's Digital Forensics Research Lab and EquisLabs noted that not only was Spanish-language disinformation the greatest foothold for RT, but also that it presents an opportunity for information to flow into the United States via its Spanish-speaking populations.[19] Understanding how other countries are targeted and the flow of content across national ecosystems is critical to understanding how demographic groups may be differentially impacted by information manipulation.

SCSP suggests the Information Integrity R&D Working Group prioritize research about the relations between national information ecosystems, with emphasis on vulnerabilities and points of fracture. Such research would develop lessons of what types of targeted activities to anticipate and various ways to respond, in addition to revealing demographic groups that are being exploited for their ability to propagate between nations.

2. Expanding Capabilities and Research Beyond Anglocentric Datasets

Recommendation: Build capacity and increase research around information manipulation targeting non-English information ecosystems.

Outcome: Broaden common understanding between private companies, government, and academia regarding the information ecosystem in other languages to better understand the differential impact of targeted influence operations.

The majority of the Internet (the interconnected series of networks) is in English, and, therefore, the information available on the Internet is primarily in English.[20] Similarly, the majority of tools and research into information manipulation and responses into influence operations is done on English-language datasets, hence English-speaking populations.[21] However, Facebook internal documents revealed that while there were some difficulties with English misinformation, the challenges were worse in other languages due to limited capacities to identify and remove misinformation in non-English information spaces.[22]

Research is slowly expanding to address other languages. For example, research indicates potential election influence and information manipulation attempting to

[18] David Klepper & Amanda Seitz, Russia Aims Ukraine Disinformation at Spanish Speakers, AP (2022).
[19] Id.
[20] Govind Bhutada, Visualizing the Most Used Languages on the Internet, Visual Capitalist (2021).
[21] See e.g., Gordon Pennycook & David G. Rand, Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality, PNAS (2019).
[22] Isabel Debre & Fares Akram, Facebook's Language Gaps Weaken Screening of Hate, Terrorism, AP (2021).

target Spanish speakers in the 2022 election cycle this year.[23] However, the proportion of research done regarding the information credibility and integrity of stories shared in other languages is disproportionately smaller than that done in English.[24]

Information ecosystems operate differently depending on the language. It is not only a matter of linguistics, but also cultural understanding of information and different primary forms of communication. The most popular global messaging apps around the world vary, with an estimated 2 billion WhatsApp, 1.2 billion WeChat, and 988 million Facebook Messenger active monthly users as of January 2022.[25] Some apps tend to be more popular in different regions or among different communities than others, such as WeChat among the Chinese-speaking community. Understanding the flow of information central to communication among different language-speaking and demographic groups helps to determine how those groups are differentially impacted.

Key research into differential impact should consider how different mitigation approaches are interpreted by users, and which ones are most impactful on a given platform. SCSP recommends the Information Integrity R&D Working Group prioritize building capacity and research beyond Anglocentric datasets and information ecosystems to create a common understanding across government, academia, and the private sector regarding how different language-speaking and demographic groups are differentially impacted by information integrity (or the lack thereof). The research would ideally also contribute to strategies that use uniquely designed tools and techniques for building resilience among especially vulnerable communities.

[23] Amanda Seitz & Will Weissert, Inside the 'Big Wave' of Misinformation Targeted at Latinos, AP (2021).
[24] Preliminary research can be found in the Sarah Nguyễn, et al., Studying Mis- and Disinformation in Asian Diasporic Communities: The Need for Critical Transnational Research beyond Anglocentrism, Harvard Kennedy School Misinformation Review (2022).
[25] Most Popular Global Mobile Messaging Apps 2022, Statistica Research Department (2022).