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

The NITRD NCO and the NSF, as part of an interagency working group on information integrity, request input from interested parties on a range of questions pertaining to Federal priorities for research and development efforts to address misinformation and disinformation. The purpose of this RFI is to understand ways in which the Federal Government might enable research and development activities to advance the trustworthiness of information, mitigate the effects of information manipulation, and foster an environment of trust and resilience in which individuals can be discerning consumers of information.

The RFI was published in the Federal Register on March 17, 2022, and the comment period was open through May 15, 2022. This document contains responses received from academia, the private sector, and civil society.

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.

Contents

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

Academic Researcher from Boston University

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.

From: Academic Researcher, **Boston University** College of Communication

BOSTON UNIVERSITY

Re: RFI Response: Information Integrity R&D

This letter is in response to the request for information from the NIRTD, NCO, and NSF regarding Federal priorities for research and development efforts to address misinformation and disinformation.

As a researcher at Boston University, I have published extensively on the topic of misinformation – a broad term that also encompasses disinformation. Drawing upon my expertise on this topic, I am pleased to address the ways the federal government might enable research and development activities to advance a) trustworthiness of information, b) mitigate effects of information manipulation, and c) foster an environment of trust in which individuals can be discerning consumers of information.

Regarding question 2 of the RFI, preserving information integrity and mitigating the effects of information manipulation, I have previously written the following with co-author Chris J. Vargo from University of Colorado, Boulder that specifically addresses key barriers for conducting information integrity research and development:

> Social media platforms, including Facebook, have entered into agreements with third parties to provide fact-checks of content circulating on their platforms. Despite having partners around the world (Goldshlager, 2020), misinformation continues (Robertson, 2020). Fact-checking partners **don't know how well their efforts perform at reducing the spread of misinformation** (Lu, 2019). Our dream research, consequently, centers around the transparency and accountability of social media efforts to address misinformation. **We need an API [application programming interface] endpoint that shows the specific actions platforms take once a message is identified as containing misinformation, including removal, warning labels, and downranking.** When considering downranking or shadow banning, even more unknowns exist. **Who still sees downranked content? How does that vary across demographics and psychographics? How do mitigation tactics affect the way audiences respond (liking, sharing, commenting, etc.)?** Researchers need visibility into these actions to assess how political ideology, media use, and media literacy interact with the steps platforms are taking to correct misinformation. Furthermore, content on social media is narrowly targeted to specific audiences. Both political and commercial ads are targeted to users

> based on their pre-existing attitudes, beliefs, and fears (Borden King, 2020; Young & McGregor, 2020). While Facebook and Twitter have robust APIs, there is no way for researchers to identify ads in real-time. We also desire the ability to assess the damage targeted influence has on platforms and believe that researchers and platforms can work together to understand these consequences and ultimately build better systems.

Moreover, regarding question 1 and understanding the media ecosystem: it is not just social media platforms with which we need to be concerned. Corporate and political actors have leveraged mainstream news media to shape citizens' views of public issues – such as climate change – through advertising. An especially deceptive form of digital content that has become ascendent over the last decade is called "native advertising," a form of sponsored content (an informational video can be found here). These native ads mimic the format of news articles and are common in nearly all US legacy news outlets including *The New York Times*, *The Wall Street Journal*, and *The Washington Post*. My research (as well as that of many other academics) has made clear that most readers do not recognize the difference between the paid native ads and genuine journalistic articles. In fact, a native advertisement from ExxonMobil that ran in *The New York Times* – and was created by their T Brand Studio – is an exhibit in a lawsuit against the fossil fuel company brought by the Massachusetts Attorney General's Office for deceptive advertising claims about climate change. While Boston University is providing focused research grants to study this type of covert disinformation, the government should be funding more research that investigates the nature and extent of this sort of practice, as well.

Pertaining to question 3, information awareness and education: media consumers are not equally influenced by deceptive content. My research indicates people who are older and less educated have more difficulty identifying online disinformation. However, my research also shows that those who are more news media literate – understanding news media operations and procedures – are more likely to identify disinformation efforts and are less likely to amplify it online. Thus, government should also be funding media literacy education efforts. A good example of this is the recent appropriation from Congress to fund a task force within the Institute of Museum and Library Services (IMLS) to develop guidance, instructional materials, and a national strategy on information literacy (page 142, here: https://appropriations.house.gov/sites/democrats.appropriations.house.gov/files/BILLS-117RCP35-JES-DIVISION-H.pdf). This national strategy should include media literacy education programs for local communities through public K-12 schools, higher education as well as public libraries and even post offices.

Thank you for the opportunity to share my expertise on this important issue. I am happy to continue this dialogue in the future.

Sincerely,
Academic Researcher from Boston University

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

Applied Research Laboratories, The University of Texas at Austin (ARL-UT)

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.

APPLIED RESEARCH LABORATORIES
The University of Texas at Austin

Response of ARL:UT to the NITRD NCO and NSF's RFI on Information Integrity R&D

Applied Research Laboratories, The University of Texas at Austin (ARL:UT)

May 12, 2022

Introduction

This document contains answers to several of the RFI questions based on our past work on information integrity. Our group, the Center for Content Understanding (CCU), is part of Applied Research Laboratories at the University of Texas at Austin (ARL:UT). ARL:UT is a Navy University Affiliated Research Center with a long history of work in acoustics, electromagnetics, and information technology. The lab works on a wide range of problems involving machine learning, including natural language processing, cybersecurity, image/video segmentation, and digital signal processing.

CCU works on various applications of machine learning and AI for the US Government, with a focus at the intersection of automated content understanding and security applications. We have completed and/or are currently working on a number of research projects related to the RFI. Our team consists of an experienced group of researchers with expertise in fabricated text detection, topic modeling, ontology-based decision support, taxonomy induction, concept extraction, transfer learning, named entity recognition, and human-guided machine learning, with many applications of these technologies in cybersecurity contexts. We have also established avenues for collaboration with external researchers in areas related to the RFI. Our team was one of four partners working with OpenAI on the detection and mitigation of malicious uses of large language models such as GPT-2. We also maintain connections with external researchers that work in areas including information integrity and the detection of disinformation campaigns.

Below are answers to the RFI questions which are most relevant to ARL:UT's previous and ongoing information integrity research. In particular, we focus on answers to questions 1, 2, and 4.

Responses to RFI Questions

Question 1: Understanding the information ecosystem

> "There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?"

Ecosystems for the creation and disseminations of propaganda and disinformation have emerged in most countries.[1] A complete understanding of these information ecosystems should include the relevant actors (such as governments, social media platforms, political parties, and marketing

[1] Bradshaw, Samantha, Hannah Bailey, and P. Howard. "Industrialized disinformation: 2020 global inventory of organized social media manipulation. Computational Propaganda Research Project." (2021).

firms[2]), and their incentives, capabilities and strategies. However, we shall focus our answer on foreign state-sponsored disinformation campaigns, since these present one of the major threats to U.S. national security, economic prosperity, and individual well-being. We describe three research challenges that will be faced when characterizing–and detecting–state-sponsored disinformation campaigns.

One of the main research challenges in understanding information manipulation is that it is a moving target. Malicious actors will modify their tactics in response to policy and technological changes. Specifically, any (human and/or automated) system for the detection of disinformation operations should also be designed for and tested under adversarial conditions, i.e., under the assumption that the disinformation agents will adapt to evade detection. Techniques from cybersecurity (red teaming and adversary emulation) could be adapted to the information integrity space to help understand the effectiveness of any proposed solutions.

A related challenge for the detection of disinformation campaigns is that automated, machine-learning based approaches are generally known to degrade as they are applied in *cross-domain* settings–i.e., when conditions differ too much from their training data. For example, we found that a classifier of state-sponsored social media accounts trained on Russian tweets (IRA 2016 campaign) suffered a large drop in accuracy when tested on tweets from the Iranian Endless Mayfly operation. More work needs to be done on ensuring that automated detectors continue to perform well on new disinformation campaigns. ARL:UT has expertise in designing and evaluating machine learning systems in cross-domain settings for various natural language processing applications such as entity recognition[3] and detection of computer-generated text[4]. We believe these sorts of techniques should be studied in the context of the use of automated approaches for detecting disinformation campaigns.

Another key research challenges in understanding state-sponsored information operations is that much of the relevant data is locked behind proprietary platforms. However, many information operations also make use of proxy websites to amplify a narrative.[5] Hyperlink data is readily available, and the analysis of hyperlink networks has shown promise in characterizing malicious (misinformational) sites[6]. ARL:UT's CCU has ongoing work in collaboration with UT Austin's Global (Dis)Information Lab (GDIL) in characterizing state-sponsored disinformation campaigns

[2] Martin, D., J. Shapiro, and J. Ilhardt. "Trends in online influence efforts." Empirical Studies of Conflict Project (2020).

[3] Rodriguez, Juan Diego, Adam Caldwell, and Alex Liu. "Transfer learning for entity recognition of novel classes." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

[4] Rodriguez, Juan Diego, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[5] Global Engagement Center, "Pillars of Russia's disinformation and propaganda ecosystem." U.S. Department of State, Washington, D.C. (2020).

[6] Sehgal, Vibhor, et al. "Mutual hyperlinking among misinformation peddlers." arXiv preprint arXiv:2104.11694 (2021).

and identifying disinformation outlets and high-risk affinity groups from freely available information on the web. If automated disinformation detection methods are accurate enough, they could be used to help prevent the narratives from a disinformation operation from spreading too widely. In addition, using link information has some potential advantages compared to an approach that only uses the document text (e.g., more easily handling multilingual disinformation campaigns). To the best of our knowledge, most research on disinformation has focused on data on proprietary platforms (e.g., social media) or on data centered around analysis of text content. Alternate approaches would be an interesting and potentially useful area of research.

Question 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. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?"

Given the complexity and magnitude of the problem, we believe that efforts to mitigate the effects of information manipulations will involve both technological as well as educational and policy initiatives. However, we will focus our answer here on the technological solutions to detecting information manipulation, as well as their shortcomings which should be addressed by future research.

While much disinformation is still generated by human troll farms, we believe it is likely to become increasingly automated as the AI technology for creating realistic content ("deepfakes") improves. The capabilities for creating synthetic video, audio, images and text has improved tremendously in the past few years and is becoming increasingly accessible. There are already examples of these technologies being used for malicious purposes. According to the Director of the U.S. National Counterintelligence and Security Center (NCSC), fake social media profiles with computer-generated faces have been used to recruit spies.[7] Even the mere possibility of a piece of content being a deepfake can have dangerous implications. For example, Gabon's military justified its attempted coup in part by claiming that a video of Gabon's current president was a deepfake.

Much research has been done on detecting deepfake images and video. Given that text is just as important a medium as images or video (e.g., for communicating online, for the public

[7] https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d

record, and for the scientific enterprise), we believe more research is needed for the detection of computer-generated text. Studies have shown that people have difficulty in distinguishing real from generated text[8,9], and tools for automated detection could help defend against computer-generated disinformation. While automated detectors of computer-generated text have been built, they can degrade when faced with text which is statistically different from the text used in building the detector.[10,11,12,13]

Our group investigated whether it is possible to adapt a detector of computer generated text for scientific articles in one technical subject area to another[14]. However, much remains to be done in ensuring that detectors are reliable and robust. Current challenges include defending against perturbed versions of generated text (adversarial examples), ensuring that detectors still work with higher quality fakes (e.g., text from more powerful text generators such as GPT-3) and detecting other forms of document tampering. For example, we found that detecting paragraph replacements is possible, but further work is required to accurately detect word or phrase replacements.[15]

Finally, in another CCU project we investigated whether it was possible to watermark text by hiding a small "message" (watermark) in the text itself, rather than in the file holding the text. We devised a method to do so such that the watermark has the desirable properties of robustness (i.e., small manipulations to the text should not remove the watermark) and imperceptibility (the watermark should look similar to the original text and the watermark should not be obvious). We believe that text-based watermarking is another promising approach that could be used to verify the information integrity of text.

[8] Kreps, Sarah, R. Miles McCain, and Miles Brundage. "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation." Journal of Experimental Political Science 9.1 (2022): 104-117.

[9] Clark, Elizabeth, et al. "All That's 'Human'Is Not Gold: Evaluating Human Evaluation of Generated Text." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

[10] Solaiman, Irene, et al. "Release strategies and the social impacts of language models." arXiv preprint arXiv:1908.09203 (2019).

[11] Ippolito, Daphne, et al. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[12] Bakhtin, Anton, et al. "Real or fake? learning to discriminate machine from human generated text." arXiv preprint arXiv:1906.03351 (2019).

[13] Stiff, Harald, and Fredrik Johansson. "Detecting computer-generated disinformation." International Journal of Data Science and Analytics (2021): 1-21.

[14] Rodriguez, Juan Diego, Todd Hay, David Gros, Zain Shamsi and Ravi Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text", Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[15] Schuster, Tal, et al. "The Limitations of Stylometry for Detecting Machine-Generated Fake News." Computational Linguistics 46.2 (2020): 499-510.

Question 4. Barriers to research

> "Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?"

One significant barrier to researchers working on automated techniques to detect disinformation campaigns is the availability of suitable data. We strongly believe that making datasets available to researchers will help shape and drive research on detecting disinformation campaigns. In particular, we believe a concentrated effort on creating data should address the following concerns:

- The datasets should specifically be tied to mitigation techniques. Several recent papers have proposed various disinformation kill chains (e.g., the MITRE disinformation kill chain included in the October 2019 report "Combatting Targeted Disinformation Campaigns" funded by the ODNI and DHS). However, many widely available datasets on detecting disinformation used in machine learning studies seem to have been created without such a kill chain or specific mitigation techniques in mind. In particular, we believe many datasets are created simply because the data is relatively straightforward to gather and/or label, not because the dataset was designed with specific, deployable methods for mitigating a disinformation campaign. Creating the dataset specifically to tie into and disrupt a particular step in a disinformation kill chain would help machine learning researchers with problem formulation and algorithmic development such that the proposed approaches could more easily be deployed in practice.
- Similarly, the datasets should be purposefully designed to mirror how detection would work in practice. For example, if detecting disinformation campaigns based on the text content of the disinformation is useful to disrupt a campaign in practice, then the focus should be on creating these types of datasets. However, it is not clear if this is the case, and datasets based on other information (e.g., network effects) should be created purposefully to enable government and industry to detect and disable future disinformation campaigns.
- Crafting a good dataset can also shape future research. For example, the types of tactics, techniques, and procedures (TTPs) used in a disinformation campaign differ based on numerous factors. A benchmark dataset will only be able to capture some of these TTPs, but making the dataset widely available will naturally attract many researchers towards detection techniques for those specific TTPs. Moreover, carefully creating a separate training set and test set for automated approaches can be used to motivate techniques that can generalize across TTPs and threat actors. Similarly, it is unclear how much data would actually be available to detect emerging disinformation campaigns in actual problem settings. The amount of data alone would shape the types of machine learning approaches studied in the academic space.

Closing Comments

Solutions to mitigate the effects of disinformation should be multifaceted; in this response we have focused on the aspects of the problem for which technological approaches could play a role. R&D from academia, UARCS and FFRDCs can be useful both for understanding the disinformation ecosystem and for developing tools to discover and disrupt future information operations. However, there are a number of challenges which will prevent research efforts from being maximually beneficial when deployed:

1. Malicious actors will modify their strategies as the environment changes (in particular, to evade detection or disruption). Systems should be designed and evaluated under adversarial conditions.
2. Further effort is required to ensure that approaches continue to work effectively with novel disinformation campaigns or novel actors.
3. The availability of suitable data to develop and test mitigation techniques is essential. Effort must be taken to ensure that such data should mirror how a disinformation campaign could be disrupted in practice.
4. Deepfakes pose a major threat to information integrity. Further work is required to develop effective and robust detectors of deepfake content, or to verify the integrity of a given piece of content in other ways.

For additional information, please feel free to contact us at:

- Center for Content Understanding
- https://www.arlut.utexas.edu/ccu/index.shtml

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

Cliff B.

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.

From: CLIFFORD B.
To: IIRD-RFI
Subject: Request for Information on Federal Priorities for Information Integrity Research and Development
Date: Tuesday, April 12, 2022 4:30:05 PM
Attachments:

I am the author of numerous information technology books, including High-Assurance Design.

I have two recommendations:

1. Social media algorithms should be off by default.

Social media amplifies disinformation by prioritizing messages that incite emotion in those who receive them. That has the effect of creating bubbles and amplifying extreme messages over thoughtful ones.

The remedy is to require that such algorithms that “choose” what we see should be off or neutral by default; a user should have to request to see algorithmically slanted information, e.g., by clicking a button that says “Show me what YOU think I would like to see”. Users who do not click that should all see the same thing.

2. Trust framework for the Internet.

People on the Internet are anonymous. As a result, if a website wants to know who someone is, that person must create an account with the website. Instead, it should be possible for someone to establish their identity with an identity provider. Then, when visiting a website, the user can choose to reveal their identity. The identity provider validates the user as being the claimed person.

The protocols for this should be standard. Today there are services that do this, but they are commercially motivated. There need to be pure identity provider services. That would put users in control of when they reveal their identity, and the identity system would be robust.

Very best,

Cliff B.

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

Coalition for Content Provenance and Authenticity (C2PA)

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.

VIA ELECTRONIC FILING

Re: C2PA RFI Response: Information Integrity R&D (Docket Number NSF - 2022-05683)

The **Coalition for Content Provenance & Authenticity (C2PA)** appreciates the opportunity to respond to the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) and National Science Foundation (NSF) Request for Information (RFI) on Federal Priorities for Information Integrity Research and Development. The Coalition for Content Provenance and Authenticity (C2PA) is an open standards organization created by members of the Content Authenticity Initiative (CAI), Project Origin, and others through the Joint Development Foundation under the Linux Foundation. The C2PA is collectively building an open technical standard to provide provenance and attribution for all forms of digital media. C2PA envisions these open standards will be adopted by content publishers, authors, and creators to build trust in the information ecosystem and ensure interoperability across the internet.

The C2PA is encouraged that NITRD NCO and NSF have dedicated time and resources to understanding ways in which the Federal Government might enable research and development activities to advance the trustworthiness of information, mitigate the effects of information manipulation, and foster an environment of trust and resilience in which individuals can better discern information online. While we are sure NSF has identified a number of potential areas for funding, C2PA recommends NSF dedicate funds to prioritize putting into the public's hands tools for determining whether content has been altered or manipulated, or is entirely synthetic, as well as educational tools for utilization of those digital and media provenance tools. We believe investment in fundamental research around the adoption of digital and media provenance standards will help to build trust and restore trust online.

1. *Understanding the information ecosystem:* There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, in addition to other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

C2PA Response

Research on visual information manipulation and the necessary trust indicators for digital content provenance are symbiotic and incredibly helpful to consumers of online information. Such research has been a challenge for the Coalition for Content Provenance and Authenticity (C2PA) due to the organization's limited bandwidth. We propose that the National Science Foundation (NSF) consider funding such research in two separate phases that each deserve time and attention. We propose that NSF distribute resources out in two phases with the results of the first round of research informing the second round of research grants.

- **Scale:** First, NSF should fund research on the prevalence of digital content fraud - specifically content created with the intent to visually deceive. This remains a significant challenge today. For content consumers, is there a quantifiable benchmark which will help explain how trustworthy digital content is? How does a consumer's level of trust in a piece of content affect their perception and understanding of it, and how does this translate to behavioral decision-making?

  Similarly, for social media platforms or browsers, what percentage of digital content inside the US is manipulated and used on their platforms for the purposes of deception or disinformation? For private industry, what is the dollar amount of visual content deception and fraud currently affecting their business or end-users? Is it possible to establish what software/AI has been used to create this content? How has the user acquired this software?

  This type of research is nuanced and challenging because it places *visual* deception at the heart of the study, rather than general disinformation, bots, or fraud studies, which are all encompassing. It remains unclear to many just how prevalent image deception is in personal, financial, national security and trust transactions taking place online. Quantifiable benchmarks could help increase awareness and preparation for countermeasures to counter this deception across society, governments, and private industry.

- **Provenance Indicators:** Building on the first proposed round of research, another research challenge for the C2PA is how to present provenance content to reassure those interacting with it online. What trust indicators would enable content consumers to accurately distinguish authenticated from standard unvetted digital content? Like the Secure Sockets Layer (SSL) in most browsers today, design affordances and UX indicators may provide trust signals to consumers of authenticated content.

  From an efficacy perspective, what visual indicators would best communicate the relevant information to consumers in a manner that they are willing and able to interpret? How do consumers respond to these indicators: do they exhibit any behavioral change when indicators are present? Are they able to accurately interpret them? How do they react to having access to this information? How do we ensure provenance indicators avoid being interpreted as a binary judgment on the content itself? What do consumers need/want to know about “truthful” content compared to content intended to deceive? Such research will help inform larger educational efforts to raise awareness of digital content provenance and how to interact with it online.

- **Privacy:** Another area of research NSF could focus on is ensuring that any digital content provenance open standard can be leveraged in ways that respect privacy and personal control of data. How willing are creators to share provenance data and how do they understand the provenance standard to make use of the data they share?

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 detecting information manipulation, discerning the influence mechanisms and the targets of the influence activities, mitigating information manipulation, assessing how individuals and organizations are likely to respond, and building resiliency against information manipulation. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?

C2PA Response

Research should focus on analyzing what demographic groups would leverage digital provenance to mitigate information manipulation. This research could also look at what factors would limit different demographic groups from leveraging digital content provenance, as well as examining how to make provenance accessible to lower-income, global majority groups. How can we avoid creating power imbalances based on provenance accessibility to creators and consumers?

3. *Information awareness and education:* A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

C2PA Response

- **Education/Awareness:** The C2PA and its members recognize that raising awareness and educating the public on digital content provenance is challenging. Most audiences expect binary decision making on existing digital content, which is not possible in real-time and/or scale. Educating on the utility of a new digital content ecosystem starting with capture to editing to display is nuanced and can involve complex concepts. Identifying the best and most feasible ways to educate society, business, government, and other facets of society on its utility will be paramount, and is, therefore, an important subject that merits substantial research funding.

  Equally important will be education and awareness campaigns on what digital content provenance does not do; (*i.e.*, binary outcomes) so as to mitigate misuse and misunderstanding of the technology and approach. The C2PA envisions various levels of educational programs targeting the many facets of the digital ecosystem. We envision three different audiences, and we recommend research that will identify best how to educate each of these three audiences:

  - *Content Consumers:* Arguably the most important audience, how can we best educate the general public on how to interact with digital content provenance when viewed online. Why it is helpful and what should be considered before making a decision of consequence based on digital content. What does the marker indicate, how to read it, and what to infer (or not) from the technology.

  - *Government*: What are the benefits of digital content provenance? How can it be used to drive increased trust, reduced fraud, and other desirable outcomes for the health of our online ecosystem. What are the threats, misuse possibilities and how can the government help to educate the public? Further, how can digital content provenance undergird the future of trust in the growing digital economy and ecosystem?
    - Government education will also need to include federal, state, and local initiatives that can be combined with existing media and online literacy efforts across the country.

  - *Business*: The private sector will need similar, but customized messaging to explain why they will need to engage, understand and adopt digital content provenance standards. While digital marketplaces and social media platforms will need to adopt the standard to help protect consumers from fraud, other industries will be able to leverage digital content provenance to deliver sustainable, accelerated, and greener services to clients and customers through trusted digital transformation. We must clarify the gaps in our understanding of business stakehlolders through research.

4. *Barriers for research:* Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that require the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

C2PA Response

- As noted earlier in the response, the C2PA lacks the resources and bandwidth to conduct the research necessary to drive educational and awareness raising at scale on digital content provenance. Furthermore, research funded directly by C2PA and its members may be regarded as biased or self-serving. Our barriers are similar to those of others in this field and we think that NSF research funding grants can help researchers overcome these challenges. Specifically, we assess this will require experienced research institutions with a track record of developing efficient and effective programs to help inform the rollout of new technology and standards. Also, the research is highly interdisciplinary and will require the involvement of researchers across various departments. Access to data will also be critical and has been a notable challenge to our initial efforts. It will be difficult to assess the prevalence of visual deception on social media platforms, peer-to-peer marketplaces, and other popular online sites without access to company data.

5. *Transition to practice:* How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

C2PA Response

We agree that the rapid transfer of information integrity into practice will be critical. Even as proof of concept to demonstrate the future of high-integrity digital content, the NSF can help the United States Federal government become one of the first implementers of the C2PA open standard. We envision a variety of federal government agencies, programs, or agency bureaus adopting the open standard and serving constituents' digital content provenance media for consumption, information, and decision making. Agencies that often interact with the general public could issue communications or media with official content provenance standards, which can be viewed on a site with the adopted standard. The following agencies could also implement digital content provenance and/or empower the public to submit provenance-enabled information as part of their online experience:

- Internal Revenue Service (IRS)
- United States Patent and Trademark Office
- Department of Commerce
- Department of State (Passport, Visas, etc.)
- Department of Homeland Security
- Department of Motor Vehicles (State)

Implementing pilot programs at these agencies with NSF support would not only prove the concept, and familiarize millions of Americans with these tools but dramatically increase the trust in digital content and ease interaction with federal agencies or state governments. For example, the U.S. Government Printing Office (GPO) has been publishing provenance-enabled (certified) PDFs on their site for almost two decades.

6. *Relevant activities:* What other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

C2PA Response

The NSF could fund different research projects studying how users' perceptions of media content change upon interacting with the different C2PA digital content provenance levels of information disclosure.

7. *Support for technological advancement:* How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission:

C2PA Response

We believe that information integrity and the adoption of media and digital provenance standards directly supports the White House OSTP's mission because it:

- Ensures the United States leads the world in technologies that are critical to our economic prosperity and national security; and
- Maintains the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values;
- Further, as detailed in the recent Declaration for the Future of the Internet, the White House, in coordination with 60 countries, announced their interest to protect consumers and promote transparency and authenticity online. Informational integrity is the best way to do this.

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

Computing Research Association's Computing Community Consortium (CCC)

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.

CCC
Computing Community Consortium
Catalyst

Response to RFI on Federal Priorities for Information Integrity Research and Development

Written by: Academic Researcher (Arizona State University), Academic Researcher (University of California, San Diego), Researcher (Microsoft Research), Academic Researcher (Indiana University, Bloomington) and Academic Researcher (University of Texas at Austin)

From the start, we would like to stress that there have been many efforts to qualify narratives and networks in this space, but relatively few that are identifying methods or solutions. A large amount of research is needed to address a complex set of challenges in the information integrity space. The Computing Community Consortium addressed a number of these points in a 2020 white paper "An Agenda for Disinformation Research". The paper describes a multi-disciplinary research agenda incorporating disinformation detection, education, measurements of impact, and a new common research infrastructure to combat disinformation and its effects upon the US and the world.

In this document we outline the specific challenges and research problems that we view as vital to mitigate the risks involved in mis/disinformation and work towards a more trustful information ecosystem.

1. *Understanding the information ecosystem: There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?*

In order to establish a common foundation for understanding misinformation within the information ecosystem, we must address the following challenges:

- **Developing better, more relevant models of social and psychological phenomena** — sociotechnical behavioral models are often outdated and do not include new sociotechnical systems. This phenomenon ties into the fact that current models do not take into account the dynamic, continuously developing world that the information ecosystem is housed.
- **Creating frameworks for complex dynamics** — which involve the source of mis/dis-information, the medium in which they propagate, and the population they

target. These frameworks become even harder to create when the population that they target is the medium in which it propagates. To complicate things even further, the medium in which it propagates is affected by the technological platforms as well as the relationships between people in those platforms and outside.

- **Detecting sources of disinformation across scales** — the scale of the impact is very different from everything else we deal with. As a result, we need to find new methods of detecting sources of disinformation.
- **Developing mechanisms for countering mis/dis-information** — thinking about how to scale the pace of propagation to suppress mis/dis-information.
- **Metrics** — there has been a lot of appreciating the problem, but not enough of developing solutions and discussing what goes into determining if these solutions are successful. We not only need to discuss what constitutes success but ways to measure the impacts that mis/dis-information has on society.
- **Data infrastructure** -— not only technical infrastructures and data sets, but also data infrastructure that acknowledges the context and socio-behavioral systems that impact these data systems.
- **Ethical research** — researchers must consider people's awareness of being included in studies and data sets and how this collection of data could adversely impact individuals going forward (e.g., someone being targeted by a disinformation campaign and then distributing the misinformation that is saved forever in a dataset). There have been studies that show people are not as aware as we think. We must also acknowledge that interventions prioritize a specific value set that may not translate across cultures, thus we should monitor and reflect on possible harms this may cause.

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. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?*

While disinformation is not new, confluence of technology and disinformation techniques (across media modalities - narrative, photo, video, etc.), creates an asymmetric vulnerability surface - it is significantly easier (cheaper, scalable) to create and propagate disinformation, than it is to counter it. Detecting and mitigating information manipulation at scale is a fundamentally interdisciplinary challenge.

Strategies for protecting information integrity must account for a wide variety of requirements. Some of these requirements may be in conflict with each other and will certainly evolve over time. There is a need for a framework that can distill and encode these heterogeneous requirements and help identify the trade-offs between them in a way that can be understood by all stakeholders.

Research is needed to understand impacts of both disinformation and mitigation techniques. To do that effectively requires bringing together computer scientists, psychologists, and social scientists. Susceptibility to disinformation based on various characteristics also needs to be rigorously studied. Particularly, "psychological" isn't on this list, even though psychological knowledge (i.e., how does information manipulation impact cognition, particularly in ways that lead to problematic feedback loops?) is one of the big things that we don't really know and is an extremely important piece to this puzzle.

3*. Information awareness and education: A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?*

An essential piece to overcoming the barriers that prevent public awareness of information manipulation is first understanding the impact that disinformation has on these communities. Currently, we don't fully understand the depth or the breadth dis/misinformation has on different communities. Are there actually different impacts based on demographics, or is it different impacts based on different features/properties (that happen to show up in different groups?). That is, could there be a "unified" model here? Citizens are confronted with a slew of information daily. There is no way for them to know what is true and what is not. After we understand the impact and the root of why audiences buy into dis/misinformation, we can start to reestablish trust with the community.

In a different vein, education/awareness might not be the right pathway here. There are lots of studies showing that people are still susceptible to mis/disinformation, even when they know that it is mis/disinformation. Take the example of vaccines. There was lots of fake information circulating the media and platforms about how dangerous the vaccines were, that the government was using vaccines to microchip citizens etc. People forgot about the vaccines they had already received and have no problem with. It is extremely hard to change people’s convictions once their mind is made up, no matter how accurate or reputable the evidence countering their claims are. As a result, any education tactics must implement the “teach the teacher model”. The information needs to come from an inside trusted source or they won’t believe it.

4. Barriers for research: Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

(1) Platforms closing or throttling data streams
(2) Intense politicization of the topic (which makes people less willing to work on it)
(3) Lack of local knowledge about what constitutes mis/disinformation in particular settings (e.g., when using memes that don't explicitly state the mis/dis)
(4) Platform belief/perception that they can't do anything about it legally (I think they're wrong, but they often espouse this belief)
(5) R&D on information integrity requires coupling across disciplinary boundaries beyond what is possible today due to the lack of, including but not limited to, appropriate organizations and promotion mechanisms.

5. *Transition to practice: How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?*

We need research incentives to shift from identifying mis/dis-information to measuring mid/dis-information and understanding how to deploy interventions.

6. *Relevant activities: What other research and development strategies, plans, or activities, domestic or in other countries, including in multilateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?*

One active, private sector activity that we would call to your attention is the Coalition for Content Provenance and Authenticity (C2PA, https://c2pa.org/). The C2PA is an international consortium consisting of technology providers and leading news organizations that is working to help combat the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content.

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

Growth Focused Insights and Research (GFI Research)

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.

NSF RFI

RFI Response: Information Integrity R&D

1 Cover

NSF RFI Information Integrity	
Title of Proposal:	Towards a better understanding of the information ecosystem
Prime Offeror: Growth Focused Insights and Research, LLC (GFI Research)	
Contact:	Researcher

2 Executive Summary

The ability to accurately identify and understand the information ecosystem is critical to the effective performance of government. The recent focus on misinformation, disinformation and mal-information (MDM) in the public domain has brought these issues to the forefront. What makes a practical understanding of the information ecosystem even more critical is the performance of government messages and actions during large scale public events and is critical to maintaining citizen confidence, domestic tranquility and preventing violence.

Large scale public events such as elections, natural disasters, terrorist attacks, public health emergencies or civil unrest provide myriad opportunities for MDM information, regardless of source, to undermine citizen confidence and destabilizing civil society. These techniques can also be applied to more routine governmental functions such as visiting national parks, buying flood insurance, employment in the Armed Services, etc.

Our response adapts both theoretical and applied research from private industry to better understand the information ecosystem and thwart efforts, whether intentional or not, to undermine our Nation's democratic, economic, geopolitical, public health and security systems.

3 Table of Contents

4 RFI Topic Areas

Our response addresses five of the seven topic areas and can help the Information Integrity Research and Development Interagency Working Group (IIRD IWG) better understand the information ecosystem and protect our institutions that are critical to the effective functioning of our society. Specifically, our response addresses

1. *Understanding the information ecosystem,*
2. *Preserving information integrity and mitigating the effects of information manipulation,*
3. *Information awareness and education,*
4. *Barriers for research,* and
5. *Transition to practice.*

In the interest of brevity and clarity of our response, we will not address each of these specifically, but will reference sub-elements of each throughout our response.

5 Problem Statement

The need to better understand the information ecosystem in order to address critical issues facing the country has never been more necessary. This understanding is even more vital in the age of social media, an explosion of traditional media and the rise of MDM information.

Large-scale public events are ripe for 'fake news', misinformation, bad information or malign forces to influence or dominate the public discourse, which can result in making the crisis worse or undermining trust in government. There is an opportunity for governments at all levels to provide resources to diminish the impacts of **"corrosive information."** Longstaff and Yang (2008) studied natural disasters and health emergencies, e.g. tsunami's and pandemic flu, finding that "trust" and "trusted information" were critical to a population's ability to recover from a crisis. If the government or the official who is disseminating the information has the trust of the population, then 'individuals, businesses and communities' will be able to recover more quickly.

US Federal agencies may need to create effective campaigns and narratives to combat these tactics and misinformation (Kangas 2019; Werchan 2019). When confronted with these challenges, government agencies should have the appropriate tools in their toolkit to understand and address the event. We suggest that Federal and State governments, departments and agencies adapt techniques from the private business sector to better understand and respond to these situations.

6 Suggested Research Tools

The core of our response is to create tools and methods to better understand how people in America understand, process and react to information in order to improve the detection, analysis, understanding, and mitigation of the threats posed by MDM information during large scale public events like natural disasters, public health crises, civil unrest, etc.

Specifically, we suggest that NSF fund multi-disciplinary research efforts that resemble the research process, tools and techniques used by the private sector.

The government can use these tools on a routine basis to understand the information ecosystem measure the impact of misinformation, disinformation and mal-information (MDM) on citizen confidence, offer insights into the best methods to counter the influence of 'corrosive information,' segment the population on key drivers, and test messages. The end result is to aid public officials' ability to better address public problems and "mitigate information manipulation, assess how individuals and organizations are likely to respond, and build resiliency against information manipulation."

The basic principles behind these tools, which we will call **CRISP**, are:

- **C**ore Insight- What are core insights that explain the situation?
- **R**TBs- What are the 'reasons to believe' (RTBs) the government's communications?
- **I**nformation- What critical information (or misinformation) is being used by the target of the message and how is it distributed?
- **S**egments- How does each key segment respond to the information, message or communication strategy?
- **P**roblem Solution- Did the government's approach (message) help solve the problem or make it worse?

The **CRISP** tools can be used across Federal Government agencies in an applied, practical manner to protect elections, respond to emergency situations, such as natural disasters, man-made events, public health crises, wild fires, climate change, etc. **CRISP** can

- be an early warning system of a shift in public confidence,
- detect the impact of 'fake news' and malign actor interference,
- react to quickly moving events in a timely manner,
- provide the ability to measure how information is being received by the public,
- identify the key segments that may be affected and RTBs, and
- provide the ability to quickly test messages to understand their impact.

A foundational research tool that we recommend is called **segmentation.** Segmentation is a standard technique that analytically identifies (typically) four to nine types (mindsets) of customers in systematic, meaningful ways to more effectively understand and message to them based on a specific context. The specific context can be a natural disaster, a pandemic (like COVID 19), a terrorist attack such as 9/11, climate change or foreign attempts to disrupt our elections. Nearly every major brand or company uses segmentation as a foundation to:

- ➢ Understand what their customers and potential customers want;
- ➢ Develop messages and communication strategies;
- ➢ Test messages, products and services;
- ➢ Measure success and track progress towards goals; and
- ➢ Conduct additional research by key segments.

The major assumption underlying segmentation is that psychographic (including behavioral and attitudinal) characteristics are, when combined with demographics, much stronger predictors of consumer preference subsequently improving communication, product development, customer service, brand equity and customer satisfaction.

Marketing communication campaigns can influence human behavior; it is especially effective when following the segmentation, targeting and positioning process (Aaker 1991). These lessons from private industry can inform policy makers on how to measure, understand and create more effective messages by applying the process of marketing communications campaigns. Government communication campaigns should be targeted and positioned correctly for each segment of the population and developed using the best practices of microeconomics, communication science and market research (Kotler & Keller 2005; Tynan & Drayton 1987). This technique is used by nearly every sector of private industry to effectively market products and services to their customers in order to influence desired outcomes (Lynn 2011.)

We recommend that an agency conduct a psychographic segmentation of the consumers and users of their information and potential targets of their messages. For instance, did the CDC or NIH have an understanding of the different types of mindsets in the U.S. when unveiling their campaigns concerning COVID? Did they test the 'brand equity' of their spokespeople and create messages accordingly? Or did they offer one primary message to everyone assuming that everyone processed information in the same manner?

Here are some questions that a segmentation could have answered regarding the Federal Government's response and messaging surrounding COVID:

- What is the brand awareness and brand equity of the CDC? NIH? Dr. Fauci?
- What are the "mindsets" of different segments of the public and how do they evaluate proscriptive behavior from the Federal Government? Do these same individuals feel the same way towards State/Local leaders, different news organizations, doctors or other non-governmental leaders/organizations?
- Were messages tested with different segments before being released?
- Was a "Creative Brief" or a "Strategic Brief" used to create a communication strategy and messages?
- What are the reasons to believe (RTB) a specific message? Are the RTBs different or the same for key segments?
- What metrics were used to measure the effectiveness of each message? And were these success metrics determined before the message was released?

7 Example Segmentation Technical Approach and Outcomes

There are many approaches that can be taken to create a segmentation of the population. Here we provide one method of conducting a psychographic segmentation that can be used as a basis for the **CRISP** tool.

Pre-Segmentation- Typically, one begins with a brief landscape assessment, including a literature review, stakeholder interviews and, if time and budget allows, an expert workshop. We also recommend analyzing existing data sets and social media to gain initial learning. The initial learning will include a prototype hypothetical segmentation.

Segmentation- The segmentation will help us identify the different subgroups in the population that share similar beliefs, attitudes, and behaviors. These segments will be used to provide insights to create a fact-based understanding of the information ecosystem, input into strategic briefs, a framework to test messages, provide a deeper understanding of how citizens consume information, provide a framework (typing tool and personas) for ongoing measurement by segment and uncover reasons to believe (RTB's).

These segments are usually created through a representative statistically valid and reliable large-scale survey and use many or most of the statistical methods commonly used to segment populations. These include k-means, tree-based methods (hierarchical clustering, CHAID/CRT, random forests), neutral networks, and others.

Segmentation frameworks and personas are generally useful for 5-15 years and will only need to be revalidated and the size of each segment adjusted slightly, unless there is a major event or shock to the system (e.g. Sept 11th, major economic upheaval, pandemic, etc.).

The segmentation will occur in seven steps. Step **one** will be a segmentation kick-off workshop to orient stakeholders and the work team to the process and align on the timeline. It should include stakeholders, clients and other firms/consultants that create messaging for the client. Step **two** will be to conduct qualitative research to better understand consumer centric language and to create a more complete and robust segmentation survey questionnaire. The **third** step is a primary large-scale survey, (n=2,400-3,300 and 15-25 minutes in length) that is representative of the population of interest. This survey will include between 50-100 psychographic statements, relevant context questions, media usage and other attitudinal/behavioral questions that are relevant to the topic area. This survey will provide a rich data set to be used to create a shorter typing tool, the core of the tracker survey, and to understand the sources of information/ misinformation and mindset of consumers.

The **fourth** step will be a segmentation selection workshop consisting of key stakeholders. This workshop should include 8-12 participants and the goal is to align on the 'best' segmentation scheme that best balances the data, context, and application of future uses for the segments. The **fifth** step is research that brings segments to life and provides more color and context for each. The **sixth** step is an activation workshop with the same participants from the kick-off workshop, which should be approximately 20-30 stakeholders, clients and support firms/consultants. There will be several deliverables, as well as a typing tool (a short set of survey questions that can be used to create these segments in future research), a final report and a set of easy to use data files.

Typing Tool- The effective rollout and continued use of segmentation schema requires an effective "typing" or classification tool so that anyone can be classified into one of the segments without requiring them to respond to the entire battery of questions initially used to create the segments. This can at times be difficult to do effectively after the fact, a problem that worsens over time. Because of this, a suggested approach is to create the segmentation and its typing tool at the same time. In fact the effectiveness of the typing tool is one criterion used in the workshop to choose the final segmentation scheme.

Tracker- We also recommend fielding a periodic tracker (monthly, quarterly or annually) that identifies key performance indicators (KPIs) that will be used to track performance and an Early

Warning System. The typing tool is used to identify each segment in order to track the KPIs by segment. In short, the segments will become the 'data cuts' to analyze KPIs. The tracker is typically a survey using the same sampling methodology and frame as the segmentation.

Expected Outcomes:

1) An understanding of the sources of information and how 'false' information is spread across the population. This includes social media and non-traditional sources of information.
2) Develop insights to develop the strategic brief that can be used to create the messages.
3) A typology (segments) of consumers and how they react to government and non-government messages.
4) A robust set of data and insights that can be used in the creative or strategic brief as guidance for creating messages or campaigns.
5) A framework, research design, survey/moderator guide, algorithm and data base tool to quickly and accurately test messages.
6) A set of key performance indicators (KPIs) to measure consumer confidence and effectiveness of messages over time, specifically we will have key measures of trust in government and in sources of information by key segments.
7) A routine tracker that measures Key Performance Indicators (KPIs) and provides an Early Warning System. This tracker can provide measures of effectiveness and help provide guidance as to impeding threats and failures of existing programs.

8 Operational Considerations

There are many different research techniques that can be used to conduct a segmentation and for the underlying resources to apply **CRISP**. We suggest that sponsoring organizations provide more direction in terms of the problem statement and expected outcomes, rather than specifics of the data collection techniques to be used.

The narrative used for the RFI is a good example of providing a statement of the problem and the solution(s) sought, without proscribing the data collection or analytical techniques to be used. This approach allows both a Program Director and a Research Vendor sufficient flexibility to respond to changing circumstances, government needs and learning based on more knowledge of the specific context.

9 Applications

The majority of our response dealt with applying these techniques to public events, usually in the context of a crises or potential to cause large scale damage. However, the **CRISP** tool and segmentation can be applied to nearly all situations, contexts, and government agencies that have a need to understand the information awareness of the public, create messages for the public or conduct education to inform the public. For instance, **CRISP**, and particularly a segmentation, can be used to create communication strategies and messages about more routine activities such as filing taxes, applying for social security, accessing VA benefits, or employment with the Federal Government.

10 Bibliography

Aaker, D., 1991, Managing Brand Equity, Free Press.

Bragg, B., 2019, Defining the Competitive Zone to Identification of Critical Capabilities, in Russian Strategic Intentions: A Strategic Multilayer Assessment Department of Defense White Paper.

Kangas, R., 2019, Recommended US Response to Russian Activities Across Central Asia, in Russian Strategic Intentions: A Strategic Multilayer Assessment Department of Defense White Paper.

Kotler, P. and Keller, K, 2005, Marketing Management, Fifteenth Edition, Prentice Hall.

Longstaff P.H., and Yang, S., 2008,Communication Management and Trust and Their Role in Building Resilience to “Surprises” such as Natural Disasters, Pandemic Flu, and Terrorism, Ecology and Society, Vol. 13, No. 1 (Jun 2008).

Lynn, M. (2011).Segmenting and targeting your market: Strategies and limitations, Cornell University, School of Hospitality Administration http://scholarship.sha.cornell.edu/articles/243.

McFaul, M. (Editor), 2019, Securing Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond, Stanford University.

Select Committee on Intelligence, 2019, Russian Active Measures Campaigns and Interference In the 2016 U.S. Election, Volume 1: Russian Efforts Against Election Infrastructure, US Senate.

Tynan, A. & Drayton, J., 1987, Market Segmentation, Journal of Marketing Management Vol 2, Issue 3.

Weitz, R., 2019, Moscow’s Gray Zone Toolkit. in Russian Strategic Intentions: A Strategic Multilayer Assessment Department of Defense White Paper.

Werchan, J., 2019, Required US Capabilities for Combatting Russian Activities Abroad in Russian Strategic Intentions: A Strategic Multilayer Assessment Department of Defense White Paper.

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

Indiana University Observatory on Social Media (IU-OSoMe)

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.

Response to NITRD/NSF Request for Information on Information Integrity

This note is a response by the Indiana University Observatory on Social Media to "Request for Information on Federal Priorities for Information Integrity Research and Development". Based on our research, we respond to the questions (highlighted in a blue italic font) from the request below.

1. *Understanding the information ecosystem:* There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

The key challenges are...

- Understanding and quantifying the key vulnerabilities of the information ecosystem. We currently know about the widespread sharing of misinformation, coordinated posting of support for minority opinions, the bias of algorithms toward popular posts, the limited attention of users, and the division of the ecosystem into echo-chamber communities or epistemic bubbles. These vulnerabilities may be due to malicious behavior but can also emerge from organic behavior on social networks and corresponding issues of platform algorithms. Further ongoing research is required to monitor these current vulnerabilities and identify new vulnerabilities.
- Understanding and quantifying how the information ecosystem is manipulated. Studies have identified how inauthentic accounts, commonly known as bots, are used to manipulate social media. They have been used to amplify specific agendas and issues, and create echo chambers. Other studies have shown how a small number of accounts are responsible for a large proportion of misinformation.
- Understanding and quantifying the real-world impacts of information manipulation. This challenge is perhaps the most important, but yet the least studied. Early work has linked online misinformation to vaccine hesitancy and impacts on democratic systems. However, causal links between misinformation at scale and offline behavior remain largely understudied. The gold standard for these causal studies involve creating interventions and studying behavioral changes.

- Developing conceptual and empirical understandings of the characteristics and mechanisms of information ecosystems. There are many disciplines working on information systems, and an organizing framework that integrates these would be of value. This would enable more holistic understanding of the vulnerabilities of the systems.

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.

Q1. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives?

The key gaps in our knowledge are as follows…

- Future-proof strategies for detection and monitoring covert manipulation of social media platforms. Covert strategies tend to use systemic flaws in social media platforms. There is an ongoing arms-race between researchers and increasingly-advanced automated or cyborg (mixed automated and manually operated) accounts. These types of accounts can use coordination [1] to give the illusion of support for a perspective, manipulate the way accounts interact with one another [2], or delete large amounts of their posts to hide the evidence [3]. Another growing concern is the manipulation of detection systems themselves.
- Strategies for detection and monitoring of overt manipulation of social media platforms. These kinds of manipulation are generally done by well-known ‘superspreader’ individuals who have large numbers of followers. Studies have linked superspreaders to widespread health-misinformation on Twitter [4].
- Developing strategies for mitigating and inoculating against the effects of manipulation. These strategies require experimental studies to understand and measure the efficacy of specific interventions in controlled and field environments. Interventions could include tools to teach misinformation literacy [5], to unfollow bad actors, to highlight quality sources in social feed ranking algorithms [6], and to increase the friction of online interactions to make it harder to flood the network.

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

There is a need to develop research to understand relationships between misinformation and certain demographics at scale. Large-scale observational studies have demonstrated links between misinformation and demographic or political groups [7,8] and there are likely to be many more associations to uncover. Studies should investigate likely disproportionate harms of misinformation on disadvantaged communities. Studies can target specific groups using surveys, controlled experiments, and interventions.

3. *Information awareness and education:* A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

Our early research [5] points to the benefits of educational tools that simulate social media environments and teach participants better social media literacy. These simulated environments allow teachers to present false information and bad actors to participants in a safe controlled way and develop their skills to recognise manipulation. However, preliminary results on this topic are sparse and suggest small effects. Challenges remain in large-scale assessments of different kinds of literacy interventions as well as in the development of realistic, age-appropriate, classroom-ready and ethically sound environments.

4. *Barriers for research:* Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

A critical and common barrier for research is the availability of data [9]. Many platforms do not release their publicly available data, and non-public data are very difficult to research. Moreover, making available harmful data that has been removed from platforms available to researchers would facilitate direct routes for inquiries that are urgently needed but currently closed [3]. It might be necessary to develop legal frameworks for the sharing of social-media data between platforms and researchers, and between researchers.

[9] Tackling misinformation: What researchers could do with social media data. Pasquetto, I. V.; Swire-Thompson, B.; and others HKS Misinformation Review, 1(8). 2020. http://doi.org/10.37016/mr-2020-49

5. *Transition to practice:* How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

We envisage two ways the federal government can help. First, they can provide legislation to make sure that all public data from social media platforms is made available to vetted researchers. Second, government can hold platforms accountable to maintaining their own standards for what is and isn't acceptable content to post online, and can back this up by developing national and international strategies for moderation and mitigation of harmful manipulation, and working with the platforms to develop their moderation and mitigation strategies.

6. *Relevant activities:* What other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

Misinformation knows no national boundaries so international coalition building around mitigation is key. There are growing calls for an international body that monitors and mitigates information integrity–specifically around large predictable events like pandemics, elections and wars. Such a body would be similar to international bodies we already have for conservation, nuclear monitoring, health, peace and security, or monetary cooperation.

7. *Support for technological advancement:* How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission:

- Ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security; and
- maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.

The U.S. is a major contributor to the pollution of our information ecosystem. However, it is a concern that reaches all nations around the world. Taking a leadership role in developing a resilient, transparent and open information ecosystem has the potential to bring the U.S. great prestige.

References

[1] Uncovering Coordinated Networks on Social Media: Methods and Case Studies. Pacheco, D.; Hui, P.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. In Proc. International AAAI Conference on Web and Social Media (ICWSM), volume 15, pages 455-466, 2021.
[2] The Manufacture of Political Echo Chambers by Follow Train Abuse on Twitter. Torres-Lugo, C.; Yang, K.; and Menczer, F. In Proc. Intl. AAAI Conf. on Web and Social Media (ICWSM), 2022. Forthcoming. Preprint arXiv 2010.13691
[3] Manipulating Twitter Through Deletions. Torres-Lugo, C.; Pote, M.; Nwala, A.; and Menczer, F. In Proc. 16th Intl. AAAI Conf. on Web and Social Media (ICWSM), 2022. Forthcoming. Preprint arXiv 2203.13893
[4] The COVID-19 Infodemic: Twitter versus Facebook. Yang, K.; Pierri, F.; Hui, P.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. Big Data & Society, 8(1): 1–16. 2021.
[5] Fakey: A Game Intervention to Improve News Literacy on Social Media. Micallef, N.; Avram, M.; Menczer, F.; and Patil, S. Proc. ACM Human-Computer Interaction, 5(CSCW1): 6. 2021.
[6] Political audience diversity and news reliability in algorithmic ranking. Bhadani, S.; Yamaya, S.; Flammini, A.; Menczer, F.; Ciampaglia, G. L.; and Nyhan, B. Nature Human Behaviour. 2022. http://doi.org/10.1038/s41562-021-01276-5
[7] Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. Pierri, F.; Perry, B.; DeVerna, M. R.; Yang, K.; Flammini, A.; Menczer, F.; and Bryden, J. Nature Scientific Reports, 2022. https://doi.org/10.1038/s41598-022-10070-w
[8] Neutral Bots Probe Political Bias on Social Media. Chen, W.; Pacheco, D.; Yang, K.; and Menczer, F. Nature Communications, 12: 5580. 2021.

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

International Research & Exchanges Board (IREX)

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.

Request for Information (RFI) No. 87 FR 15274

Federal Priorities for Information Integrity Research and Development

Organization Name	Point of Contact
IREX	**Researcher from IREX**

Introduction

Established in 1968, IREX is a global non-profit organization headquartered in Washington, DC. IREX has embraced a people-centered approach, investing in development that maximizes human potential and improves the conditions that help people thrive to promote positive lasting change globally. With an annual portfolio of $90 million, IREX maintains presence in over 100 countries through innovative programs that address information manipulation, empower youth, cultivate leaders, strengthen institutions, and extend access to quality education and information. This global portfolio includes the use of multiple strategies to help individuals, communities, and organizations protect information integrity and build resilience to information manipulation and produce a healthy media environment, including trainings, campaigns, institutional strengthening of media and watchdog groups among others.

IREX would like to thank Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) and National Science Foundation (NSF) for the opportunity to provide feedback related to specific questions that reflect lessons learned and identified best practices from our global program experience.

Information Requested:

1. *Understanding the information ecosystem:* There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

Research is often poorly resourced and/or siloed into a single discipline, e.g. technology, psychology, etc. when a multidisciplinary approach is required to develop a shared understanding of information manipulation. IREX believes that developing a common understanding of information manipulation requires exploration of the technological aspects of the ecosystem (e.g., bots, artificial intelligence (AI), and virtual reality (VR)) and the strengths and challenges of people who engage in the system (e.g., users, policy makers, ecosystem developers) (Glasgow et al., 2012). In an effort to explore the complexity of these interactions, IREX created the Vibrant Information Barometer (VIBE), IREX's annual index to track how information is produced, spread, consumed, and used. Another of IREX's efforts, Securing Access to Free Expression (SAFE)-L2D, explores how journalists can be supported to engage with information and information consumers securely to prevent the spread of manipulated information. Investing in research that would enable practical application of multidisciplinary approaches, such as VIBE and SAFE-L2D, will help develop meaningful and differentiated understanding of the information ecosystem. In particular, the National Science Foundation's Accelerator program should build on previous research successes related to information manipulation, issuing opportunities for multidisciplinary teams to explore systems-levels approaches to develop common terminologies, theories, and understandings of the protective and risk factors for information manipulation within an ecosystem. To reach a common understanding of information manipulation and use it to engage multidisciplinary stakeholders, we need to understand its impact in the information ecosystem across sectors.

The politicization of information manipulation research creates obstacles to achieving a common understanding of information manipulation in the information space and generalizable, actionable, and credible results. Studies suggest that manipulative information creates pernicious cycles that foster disbelief, undermining trust in credible sources of information, attacking established methods of creating a common understanding of issues and solutions, and perpetuating belief in manipulated information (Jaiswal et al., 2020; Marwick & Lewis, 2017; Sharp, 2008). IREX proposes several methods for addressing these challenges. First, IREX supports this RFI's framing of the issue as "information manipulation," as this language establishes a productive line of inquiry without using now-politicized terms, such as "disinformation, fake news, misinformation, propaganda, conspiracy." Second, given the ongoing efforts to spread manipulated information, there is a need to ensure that even the term "information manipulation" does not become politicized. The federal government may seek to promote the acceptability of information manipulation research by seeking bi-partisan agreement on the terms around information manipulation and issuing RFIs and funding opportunities that address information manipulation regardless of its source, target audience, or intended effect. Encouraging agreement around terminology and definitions may also promote a common understanding of the challenges within ecosystems, such as Youtube, Facebook and Twitter, which apply unique terminologies and rules to addressing problematic information sharing (Aral & Eckles, 2019; Geeng et al., 2020).

Online information ecosystems provide differential levels of transparency and platforms for data access (e.g., public application programming interfaces (APIs)), creating barriers to the development of a common understanding of information manipulation. Research on online platforms is essential to understanding of the ever-evolving information landscape, especially its incentives and social and psychological dimensions. Given the global reach of these platforms, research on these ecosystems have the potential to produce generalizable inferences about the effectiveness of interventions not just in high income countries, but in low and middle income countries where research is lacking. In the aftermath of the 2018 Cambridge Analytica scandal, platforms restricted researcher API access (Tromble, 2021; Vallury et al., 2021; Walker et al., 2019). Researchers' lack of API restricts access to the data required for research involving rigorous causal inference (Pfeffer et al., 2022; Tsou, 2015), inhibits implementation of studies that could lead to understanding of behaviors associated with information manipulation, creates laborintensive data collection conditions (Aral & Eckles, 2019), and limits the ability of researchers to collect "large or representative samples of real-world events" (Walker et al., 2019). Twitter's history of providing public API for researchers has made research over-reliant on Twitter-related content even though Twitter (and other platforms) remain a black box for researchers (Pfeffer et al., 2022; Tromble, 2021). Ecosystem creators' deeper engagement and collaboration with researchers would be essential to move the research field forward from weaker quasi-experiments or observational studies to causal inference with actionable results (Tromble, 2021).

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

IREX Ecosystem of Resilience Approach to Information Manipulation

Community - In L2D, we use the principles of behavior change to introduce new behaviors, teach emotional regulation, cross-checking sources, and waiting before immediately sharing content, and create a close-knit community to reinforce these new behaviors.

Media - We support locally led strategies to address online disinformation. In Mozambique, we used machine learning and our Media Content Analysis Tool to identify bias in news articles. This type of technology has the potential to strengthen news outlets' vetting processes at scale.

Information consumers - We are building plug-ins, apps, and other tools that alert users when their data is being harvested for disinformation campaigns.

To mitigate the effects of information manipulation, research should focus on the resilience ecosystem, developing and testing multi-level theoretical models. Learn to Discern (L2D), IREX's flagship initiative for empowering critical information engagement, has been adapted in 20 countries and with multiple, diverse populations ranging in age, native language, socioeconomic status, and geography. In implementing L2D and its multiple iterations including L2D-SAFE, L2D for decision makers, digital wellness, and others, we have come to view resilience as an ecosystem, equally dependent upon measures on *system, network, organization*, and *individual* levels of resilience building. Research that addresses multiple levels of behavioral determinants have the potential to be more effective than individual-level interventions alone and utilizing a social ecological approach would be critical for addressing information manipulation (Eldredge et al., 2016). There are multiple theoretical models that could be reviewed against the complexities of internal and external incentives, motivations, and specifics of information engagement upon which a multi-level model can be built including, for example, inoculation theory, (Compton et al., 2021) social cognitive theory (Bandura, 1977, 2001), information–motivation–

behavioral skills model (Fisher et al., 2003), etc. theories about decision making and rationality, such as the System I and System II thinking (Tversky & Kahneman, 1989) that have informed our L2D work for example. We have combined insights from their research with adapted elements of the information-(added skills) - motivation-behavior (added action and network to reinforce and maintain) and inoculation theory as well as best practices from initiatives and research in fighting addiction, gang membership, and other internally and externally absorptive phenomena. We are aware that we have only scratched a surface of what research could be informing truly impactful initiatives but know that all of these and many other models should also be considered, especially when trying to understand the intended and unintended abuse of our cognitive processes in today's information ecosystem.

To promote equity in information manipulation research, strong and universal ethical guidelines are needed. The study of information manipulation is burgeoning with few common guidelines for ethical conduct of such research. Ethical guidelines that center the participants in research, engage communities of interest at the outset of study development, and use human-centered techniques are needed to promote safety while enabling the study to draw rigorous causal inference. Of particular interest and concern is the use of inoculation theory, which may be effective in building resilience around information manipulation, but also requires that individuals are exposed to information manipulation as part of the intervention (Compton et al., 2021). The creation of these guidelines may be informed by other disciplines that have long grappled with the potential ethical issues, including, for example, violence prevention (Hartmann & Krishnan, 2014, 2016). In addition, interventions like this should be designed with intentions to disseminate and apply findings outside of the research environment. Because of the nature of the topic, it is unwise to rely on voluntary uptake/audiences seeping out these tools, and many of these tools, especially targeting children and young adults, such as games, require an adult champion/promoter such as educators, parents, and caregivers. These stakeholders need to be on board with using "anti-hero" and other tools that utilize negative examples. Their reluctance to use them, should it be registered, must be considered a valid obstacle to behavior change.

IREX recommends recognizing and deliberately focusing efforts on supporting locally-led strategies to build trust. The shift in the information infrastructure, which has so gravely affected the traditional media market, has enabled an unprecedented volume, speed, and reach of malign, low quality, or simply incorrect information, ranging from effective and targeted propaganda and influence campaigns, to misinformation, clickbait, and other forms of information "noise". This has drastic consequences on the trust in media and civil society; on audience's attention spans and on the ability to agree on facts; and on the business models and incentives for information producers. Most alarmingly, disinformation has been used as an effective tool to drive and aggravate divisions, populism, and polarization—undermining social cohesion, peace, and reconciliation processes, democracy, and rule of law. Trust, once lost, is hard to restore, and has significant, negative implications for democratic integrity and the media sector's watchdog role.

To maintain a healthy information environment, research on the long-term effects of interventions on attitudes, skills, and behaviors is needed. Extant literature typically focuses on knowledge, skills, and intentions that are reported during the experiment or shortly thereafter. This approach creates multiple gaps in our understanding of interventions to create and maintain a healthy information environment. First, with few exceptions (Murrock et al., 2018), we lack knowledge about if the target skills and practices are maintained to reduce vulnerability to information manipulation. In particular, cognitive behavioral and emotional regulation components as well as digital wellness have shown promise in multiple studies, but their effectiveness post-intervention require further exploration (Karduni, 2019). Second, we lack knowledge about behavioral change associated with interventions, which could provide unbiased proof of the intervention effectiveness and increase the causal inference that can be drawn by our studies. Research that supports longitudinal studies and collaboration with ecosystem creators could help fill these important gaps in knowledge.

3. Information awareness and education: A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

To create and support resilience against information manipulation, research should move beyond exploring a single factor or skill to examining the package of skills and practices required to create resistance to information manipulation. In general, studies seek to isolate the effect of singular intervention techniques to build resiliency to manipulated information (e.g., analytic thinking, cognitive reflection, warning labels, emotional regulation, digital wellness) (Pennycook & Rand, 2021). However, it is extremely unlikely that such a complex issue can be solved through one skill or by one person alone or that the same skill is responsible for resilience in all population. Information manipulation creates emotional, mental, and even physical reactions (Marwick & Lewis, 2017; Swire-Thompson & Lazer, 2020) and has roots in systems-level inequalities including, for example, structural racism and discrimination (Cooke, 2017, 2018a, 2018b). Research should leverage agile study designs, such as the MOST framework (Collins et al., 2007) to understand the package of skills and practices needed to build resilience. An equitable approach would focus on adapting and testing the mechanisms by which these packages of skills can be delivered to different groups in a responsible and effective manner, as well as obstacles and reasons for why some populations are struggling with gaining these skills – aspects of access, format, etc.

Research must move beyond the language and assumptions of partisanship as cause or vulnerability to information manipulation to explore additional, modifiable factors. Studies often cite partisanship or identification with a specific ideology as risk factors for belief in manipulated information (Allcott & Gentzkow, 2017; Geeng et al., 2020). While factors such as ideology cannot be ignored, we also recognize that manipulated information may reach and impact individuals differently based on multiple and often intersecting characteristics (e.g., socioeconomic status, experience of structural discrimination etc.) For example, marginalized groups are often targets of manipulated information rooted in historical inequities and there is a need learn more about building and sustaining behaviors around navigating hate speech and targeted disinformation (Cooke, 2021). We must also recognize the role of information engagement in shaping ideologies and belonging to partisan group – it is not inconceivable that information engagement habits that increase vulnerability to disinformation are, at least partially, the cause, and not the result of these characteristics. Research that seeks to understand what constitutes healthy resilience to information manipulation and what contributes to individuals and communities developing or not developing competencies that are associated with this resilience would have shed the light on approaches to increase it.

Public spaces and areas of community gathering could be an important educational pathway for addressing information manipulation, but not without significant investment in local resources. Related to the previous point, there are notable community-level disparities in internet access, mental and emotional health resources, educational opportunities, local new outlets/papers and these disparities may make some community members more vulnerable to manipulated information. In many instances, education on how to use the online space in a way that does not undermine democratic values and human rights (both our own and others) is quickly becoming inequitable. It is increasingly the case that well-resourced communities/countries/school districts will (and many already do) offer these skills to their members, citizens, and students, deepening divisions (Vogt & Scott, Forthcoming). Multiple solutions can be offered - whether policy and mandates, thoughtful and resonant resources, or trusted local community advocates for equitable access – as key for preventing further fracturing along the "factual" fault lines in societies worldwide. Policies that mainstream media and information literacy education, create safe public spaces for engagement using, for example, the use of urban planning and publicly owned spaces for engagement, are needed to close the growing divides and focus on especially vulnerable populations targeted with conspiracy theories and propaganda.

Greater public awareness may be raised through research on the second and third order effects of manipulated information. Understanding the cost of manipulated information on all aspects of societies, ranging from the individual to the society, may help galvanize support for interventions to address the widespread issue of manipulated information. Given the social costs of manipulated information in relation to the COVID-19 pandemic alone, the costs of investing in people and educational systems as part of primary strategies will pale in comparison.

REFERENCES

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211-236. https://doi.org/10.1257/jep.31.2.211

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858-861. https://doi.org/doi:10.1126/science.aaw8243

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, *84*(2), 191.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1-26.

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *American Journal of Preventive Medicine*, *32*(5), S112-S118.

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.

Cooke, N. A. (2017). Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The library quarterly*, *87*(3), 211-221.

Cooke, N. A. (2018a). Critical literacy as an approach to combating cultural misinformation/disinformation on the Internet. *Information literacy and libraries in the age of fake news*, 36-51.

Cooke, N. A. (2018b). *Fake news and alternative facts: Information literacy in a post-truth era*. American Library Association.

Cooke, N. A. (2021). Tell Me Sweet Little Lies: Racism as a Form of Persistent Malinformation. PIL Provocation Series. Volume 1, Number 4. *Project Information Literacy*.

Eldredge, L. K. B., Markham, C. M., Kok, G., Ruiter, R. A., & Parcel, G. S. (2016). *Planning health promotion programs: an intervention mapping approach*. John Wiley & Sons.

Fisher, W. A., Fisher, J. D., & Harman, J. (2003). The information-motivation-behavioral skills model: A general social psychological approach to understanding and promoting health behavior. *Social psychological foundations of health and illness*, *22*(4), 82-106.

Geeng, C., Yee, S., & Roesner, F. (2020). Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Glasgow, R. E., Green, L. W., Taylor, M. V., & Stange, K. C. (2012). An evidence integration triangle for aligning science with policy and practice. *American Journal of Preventive Medicine*, *42*(6), 646-654.

Hartmann, M., & Krishnan, S. (2014). *Ethical and Safety Recommendations for Intervention Research on Violence Against Women*.

Hartmann, M., & Krishnan, S. (2016). *Ethical and safety recommendations for intervention research on violence against women. Building on lessons from the WHO publication.* (Putting women first: ethical and safety recommendations for research on domestic violence against women., Issue.

Jaiswal, J., LoSchiavo, C., & Perlman, D. C. (2020). Disinformation, Misinformation and Inequality-Driven Mistrust in the Time of COVID-19: Lessons Unlearned from AIDS Denialism. *AIDS and behavior*, *24*(10), 2776-2780. https://doi.org/10.1007/s10461-020-02925-y

Karduni, A. (2019). Human-misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information. *arXiv preprint arXiv:1903.07136*.

Marwick, A. E., & Lewis, R. (2017). Media manipulation and disinformation online.

Murrock, E., Amulya, J., Druckman, M., & Liubyva, T. (2018). Winning the war on state-sponsored propaganda: Results from an impact study of a Ukrainian news media and information literacy program. *Journal of Media Literacy Education*, *10*(2), 53-85.

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388-402. https://doi.org/https://doi.org/10.1016/j.tics.2021.02.007

Pfeffer, J., Mooseder, A., Hammer, L., Stritzel, O., & Garcia, D. (2022). This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. *arXiv preprint arXiv:2204.02290*.

Sharp, D. (2008). Advances in conspiracy theory. *Lancet*, *372*(9647), 1371-1372. https://doi.org/10.1016/s0140-6736(08)61570-6

Swire-Thompson, B., & Lazer, D. (2020). Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health*, *41*, 433-451. https://doi.org/10.1146/annurev-publhealth-040119-094127

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media+ Society*, *7*(1), 2056305121988929.

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, *42*(sup1), 70-74.

Tversky, A., & Kahneman, D. (1989). Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers* (pp. 81-126). Springer.

Vallury, K. D., Baird, B., Miller, E., & Ward, P. (2021). Going Viral: Researching Safely on Social Media. *J Med Internet Res*, *23*(12), e29737. https://doi.org/10.2196/29737

Vogt, K., & Scott, S. (Forthcoming). Building individual and community resilience to manipulative information online hate – lessons learned. In B. Bahador, R. Brown, C. Hammer, & L. Livingston (Eds.), *Countering online hate and its offline consequences in conflict-fragile settings*. GWU Press.

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, *22*(11), 1531-1543. https://doi.org/10.1080/1369118X.2019.1648536

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

Jonathan M.

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.

Comment of Academic Researchers on Federal Priorities for Information Integrity Research and Development

Thank you for the opportunity to inform the Interagency Working Group's strategic plan for a whole-of-government approach to information integrity research and development.

We are academic researchers who study information integrity topics from diverse disciplinary perspectives, including communication, computer science, economics, and law. Our scholarship aims to characterize the scope, scale, and effects of mis- and disinformation.[1] We also seek to mitigate these challenges with our research, by examining methods for identifying and responding to mis- and disinformation.[2]

Before turning to our substantive suggestions for the strategic plan, we would like to offer a recommendation for the plan's framing and the goals it may articulate. While the White House announcement of the Working Group and the Working Group's RFI both discuss "understand[ing] the full information ecosystem," the documents emphasize "information manipulation" such as mis- and disinformation. We urge the Working

[1] *E.g.*, Shan Jiang et al., *Modeling and Measuring Expressed (Dis)belief in (Mis)information*, ICWSM (2020); Miriam J. Metzger et al., *From Dark to Light: The Many Shades of Sharing Misinformation Online*, Media & Communication (2021); Tanushree Mitra & Eric Gilbert, *CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations*, ICWSM (2015); Katherine Ognyanova et al., *Misinformation in Action*, HKS Misinformation Review (2020); Mattia Samory & Tanushree Mitra, *Conspiracies Online: User discussions in a Conspiracy Community Following Dramatic Events*, ICWSM (2018); Julio C.S. Reis et al., *A Dataset of Fact-checked Images Shared on WhatsApp During the Brazilian and Indian Elections*, ICWSM (2020).

[2] *E.g.*, Austin Hounsel et al., *Identifying Disinformation Websites Using Infrastructure Features*, FOCI (2020); Ben Kaiser et al., *Adapting Security Warnings to Counter Online Disinformation*, Usenix Security (2021); J. Nathan Matias, *Nudging Algorithms by Influencing Human Behavior: Effects of Encouraging Fact-Checking on News Rankings* (2020); Julio C. S. Reis et al., *Can WhatsApp Benefit from Debunked Fact-checked Stories to Reduce Misinformation?*, HKS Misinformation Review (2020).

Group to expressly address a broader class of potentially harmful information[3] and to elevate study of the overall information ecosystem in the strategic plan.[4] There is a growing body of evidence that factually false or misleading information accounts for a comparatively small component of the overall information ecosystem.[5] Politically polarized or emotionally charged information is much more prevalent and is an important contributing factor to the individual and societal harms that prompted the Working Group's formation.[6] Advancing understanding of the overall information ecosystem, including the most effective sources of information and trends in the ecosystem over time,[7] will also be essential for contextualizing information integrity challenges and developing possible responses. While our comment focuses on mis- and disinformation research, consistent with the emphasis in the White House announcement and RFI, our observations are generally applicable to study of other types of potentially harmful information and the overall information ecosystem.

We offer four recommendations for the strategic plan that the Working Group is developing. First, federal government R&D efforts related to information integrity would benefit from greater clarity of each agency's responsibilities and coordination of relevant initiatives across agencies. Second, federal research funding should continue to rapidly scale up in this topic area. Third, federal R&D strategy should prioritize enabling methodological advances in information integrity research, by building reusable technical infrastructure, addressing legal ambiguity, and facilitating access to online platform data. Fourth, federal strategy should advance the interdisciplinary research and teaching infrastructure necessary for this area of R&D.

[3] The Working Group could, for example, clarify that "information manipulation" and "manipulated information" include a broader class of potentially harmful information, such as instances of politically polarized or emotionally charged information that could cause individual or societal harm.

[4] We commend the administration for including "understand[ing] the full information ecosystem" within the Working Group's charge. Our recommendation is not a substantive change to the Working Group's scope, but rather, a rebalancing of emphasis and priorities in the strategic plan.

[5] *E.g.*, Jennifer Allen et al., *Evaluating the Fake News Problem at the Scale of the Information Ecosystem*, Science Advances (2020); Andrew Guess et al., *Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook*, Science Advances (2019)

[6] *See generally* Joshua A. Tucker et al., *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature* (2018).

[7] Trends that may be particularly relevant include the transitions from local to national news coverage, from institutional journalism to social media, from print and TV to online information sources, and from desktop to mobile engagement with information.

1. Federal research and development efforts related to information integrity would benefit from greater clarity of agency responsibility and interagency coordination.

Information integrity research defies easy categorization. Within academia, these topics typically span academic disciplines and fields within disciplines. Similarly, within the federal government, information integrity R&D crosses agency authorities, agency equities, and programs within agencies. There is currently limited clarity of what each agency's responsibilities are and which information integrity topics fall within particular agency programs. There is also limited coordination of information integrity R&D across the federal government.

This lack of clarity and coordination has significant consequences. Researchers and other experts external to the federal government may fall into silos, interacting with agencies, programs, and staff based predominantly on familiarity and preexisting professional networks rather than relevant expertise, experience, and needs. The status quo deprives the field of valuable bidirectional information exchange with the federal government and limits opportunities for federal collaboration with and support of external research.

Consider, as an example, the developing area of research on interventions to address mis- and disinformation related to COVID-19 on online platforms. Which agencies should be responsible for research efforts in the area? NSF is a logical candidate, because of its broad responsibilities related to research support. NIH and CDC also play vital roles, because the topic relates to medicine and public health. But the list continues: this type of mis- and disinformation can have implications for armed forces readiness, creating a role for the Department of Defense and its components, including DARPA. This type of content can also have foreign policy implications, or be associated with foreign information operations, such that the State Department and the Intelligence Community have relevant equities. Because this type of content can impact critical infrastructure, especially in the healthcare and emergency services sectors, the Department of Homeland Security has a role. When there is a commercial aspect to false or misleading information, or research examines online platform practices, the Federal Trade Commission may have a role. The list could go on.

These complexities continue within agencies. At NSF, as an example, this type of research may fit within the Directorate for Computer and Information Science and Engineering (CISE) or the Directorate for Social, Behavioral, and Economic Sciences (SBE), among other agency components. And within CISE, the Secure and Trustworthy Cyberspace, Human-Centered Computing, and the Designing Accountable Software Systems programs are all possible points of connection.

This fragmented R&D landscape leads to missed opportunities. Researchers who approach the topic from a computer science perspective predominantly interact with one directorate at NSF (CISE), those who come from social science perspectives interact with another directorate (SBE), and those from medical and public health backgrounds engage with entirely different agencies (NIH and CDC). Meanwhile, interaction with other relevant components of the federal government—especially those with operational responsibilities, which may have valuable experience and expertise in the problem area—is generally limited. The result is missed opportunities for information exchange, research collaboration, research support, and better informed policymaking.

As a starting point, we encourage the Working Group to develop a public directory of federal agencies and programs relevant to information integrity R&D, with a summary of responsibilities and point of contact for each.[8] This simple step could have significant benefits. Earlier this year, NSF division directors published a brief Dear Colleague letter clarifying that the Secure and Trustworthy Cyberspace program supports information integrity research and describing the topics, disciplines, and methods that are within scope.[9] The letter has been invaluable to the information integrity research community, highlighting a point of connection with the federal government that was unfamiliar to some researchers and prompting new interdisciplinary collaborations. We recommend that the Working Group consider replicating that model across federal R&D.

[8] We do not take a position on which component of the federal government should be responsible for developing the directory or hosting the convening that we propose. There are a range of possible models, such as the recent National Artificial Intelligence Initiative and the National Quantum Coordination Office. We also take no position on the right allocation of responsibility among agencies or programs—which may benefit from overlapping portfolios, because different components of the federal government have different equities, resources, and types of expertise.

[9] Sylvia Butterfield et al., NSF 22-050, *Dear Colleague Letter: Inviting Proposals Related to Information Integrity to the Secure and Trustworthy Cyberspace Program* (Feb. 24, 2022).

Another valuable step would be coordinating a recurring event where policymakers and R&D practitioners across the federal government can engage with external researchers and other experts. A potential model is the annual data privacy conference (PrivacyCon) organized by the Federal Trade Commission, which for six years has convened federal privacy regulators and leading experts to exchange ideas, inform policy, and advance the research field.[10] A similar recurring event, connecting federal officials with external experts on information integrity, would be an asset to the field.

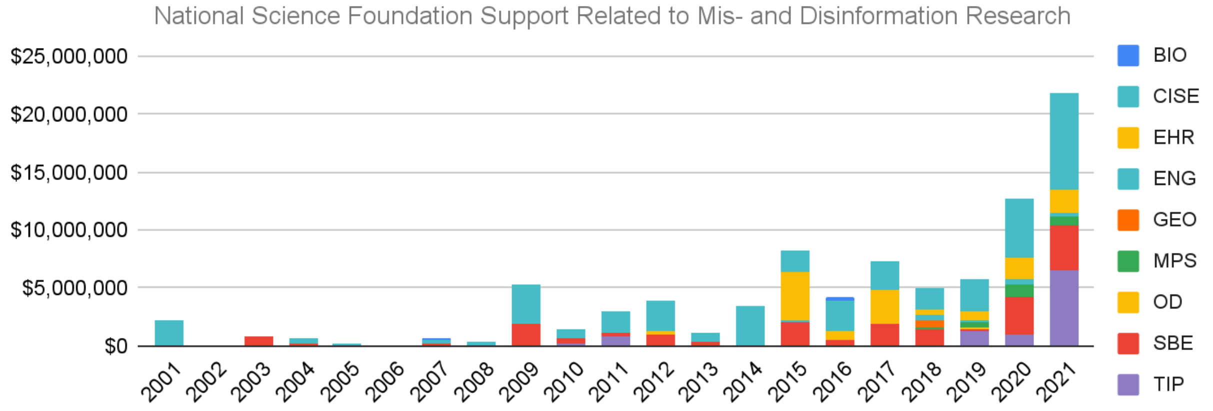
2. The federal government should continue to significantly scale up research support related to information integrity.

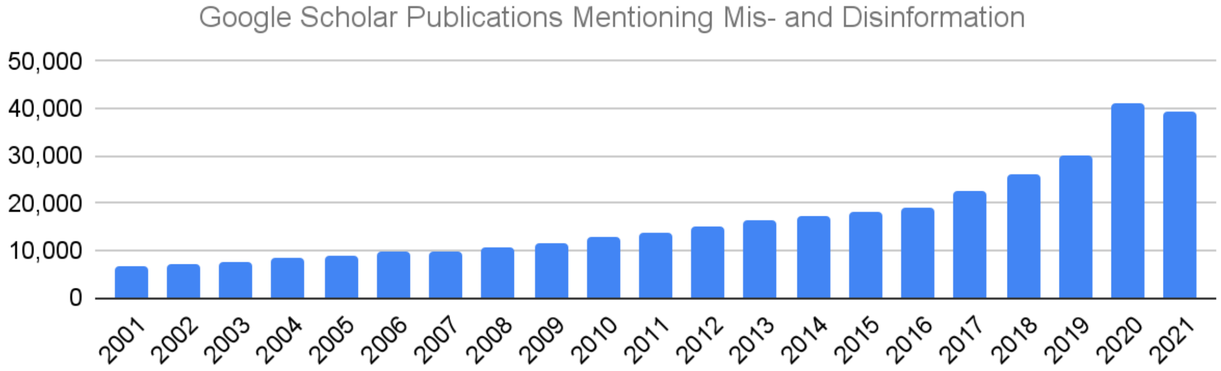
There is, at present, an extreme mismatch between the magnitude of the societal challenge posed by information integrity problems, the level and timeliness of federal support, and the rate of research productivity.

As rough empirical guides, we have compiled National Science Foundation awards that mention "misinformation" or "disinformation," as well as publications on Google Scholar that mention either of those terms.[11]

[10] FTC staff also contribute to the organization of an annual workshop on consumer protection and technology (ConPro), which is another valuable possible model for the Working Group.

[11] The NSF figure presents the amount awarded to date by award year, relying on data exported from the NSF award search website. The Google Scholar figure presents the count of papers by year of publication. The keyword matching that we use to generate data on NSF awards and Google Scholar publications is inexact. Grants related to information integrity, including a major $15.7 million award to Northeastern University, do not consistently use the terms "misinformation" or "disinformation" in their title or abstract. Publications may also use different terminology for information integrity concepts, and mentions of the two keywords is not conclusive that a grant or publication is closely related to information integrity.





A pair of trends are evident from this rudimentary data. First, the field of information integrity research has been growing rapidly, and federal support has not kept pace. While philanthropies have stepped in to augment federal funding, such as the Knight Foundation, the NetGain Partnership, and the Omidyar Network, these sources of support are not sufficient to sustain and grow an entire area of research. Corporate funding related to information integrity, meanwhile, remains relatively limited—Meta, for instance, awards about $1 million in competitive information integrity grants annually. Technology firms face complex incentives related to mis- and disinformation, and the federal government should not count on the private sector to address funding gaps or prioritize the most societally urgent aspects of information integrity.

A second readily apparent trend is that support for information integrity research began to significantly increase in 2020. We commend NSF's recent prioritization of this area, especially through the CISE, SBE, and EHR directorates and the TIP directorate's Convergence Accelerator program.

We encourage the Working Group to consider how to ensure consistent and continued growth in the level of support for this research area until the federal government reaches a level commensurate with the societal urgency of the topic. The Working Group may find it helpful to more rigorously benchmark public sector, private sector, and philanthropic support for information integrity research in comparison to other R&D priority areas. Examining past and potential barriers to federal information integrity R&D would also be a valuable step.

3. Federal R&D strategy should prioritize enabling methodological advances for information integrity research, by building reusable technical infrastructure, addressing legal ambiguity, and facilitating access to online platform data.

Research related to information integrity often uses a familiar set of methods, such as interviews, surveys, laboratory experiments, crowd tasks, web crawls, and social media archive analysis. These methods are valuable, and we use them in our own research. But these methods have significant limitations: they often are not representative of user experiences and activities, especially when users are interacting with personalized or targeted content. These methods limit the types of studies that researchers can carry out, and results from studies using these methods can have questionable ecological validity.

As an example, there are hundreds of research publications that explore how to design user interface interventions to counteract mis- and disinformation.[12] The overwhelming majority of recent papers use simple survey-like methods—asking participants to scroll a simulated social media feed and self-report perceptions and predicted actions—as well as similar warning and notice labels. Researchers in the information integrity field recognize that these methods are not very realistic and that there is a pressing need to examine alternative interventions beyond labels attached to content. But, absent methodological innovation, implementing more realistic research designs and exploring alternative interventions will continue to be elusive.

[12] *See* Laura Courchesne et al., *Review of Social Science Research on the Impact of Countermeasures Against Influence Operations*, HKS Misinformation Review (2021).

We urge the Working Group to examine how federal R&D strategy might support methodological advances for information integrity research. One important path forward is building reusable technical infrastructure, offering new data sources and capabilities for measurement and intervention experiments.[13] Mozilla Rally, for example, is an initiative that enables running research studies in participant web browsers, observing real online experiences and deploying interventions in real online settings. By making a significant upfront investment in browser instrumentation, a data analysis environment, and participant recruiting, Rally enables research with innovative methods at relatively low marginal cost and effort. Similar promising projects include The Markup's Citizen Browser (which has enabled dozens of projects examining the Facebook information ecosystem), the NYU Ad Observatory (which provides valuable ongoing data about Facebook advertising content), and Northeastern's upcoming NSF-supported Observatory for Online Human and Platform Behavior (which will enable observational study of real experiences and activities online). The federal government could have a transformative impact on the information integrity research field by supporting initiatives like these.

Another important step for enabling methodological advances in information integrity research would be addressing legal ambiguities. When researchers conduct independent study of an online platform, regrettably, sometimes the platform threatens legal action against the research team. A particularly vivid example of this issue occurred last year, when Facebook sent a cease-and-desist demand to the NYU Ad Observatory team—and then shut down the team's access to Facebook services on a privacy pretext, prompting a rebuke from the FTC. While the NYU team chose to stand up to Facebook, legal threats like these are common and create a chilling effect for independent research on online services.

The federal strategy for information integrity R&D is an opportunity to develop a whole-of-government approach toward a safe harbor for this type of independent platform research, which is vital for progress on information integrity challenges. The strategy could, for example, encourage the Department of Justice to reconsider its guidance on the Computer Fraud and Abuse Act, a notoriously ambiguous law that the

[13] *See* Elizabeth Hansen Shapiro et al., *New Approaches to Platform Data Research* (Feb. 2021).

Supreme Court recently significantly narrowed.[14] DOJ's guidance currently only addresses certain types of computer security research, and it has not updated the guidance in light of new Supreme Court precedent. Similarly, the strategy could encourage the Department of Commerce and DOJ to consider advocating for a Digital Millennium Copyright Act safe harbor for this type of research. The DMCA is another broad and vague law that could be interpreted to encompass certain information integrity research, and the Copyright Office conducts an exemption rulemaking every three years where Commerce and DOJ's views have significant weight.

A final potential path forward for methodological innovation in information integrity research is access to data held by online platforms. There have been a number of recent proposals for this approach to platform regulation, including the European Union's Digital Services Act and recent legislation in the House and Senate. We encourage the Working Group to consider how the administration can best engage with these proposals, such as by advocating for U.S. researcher access under the EU DSA or supporting particular legislation in Congress.

4. Federal R&D strategy for information integrity should prioritize building the emerging field of interdisciplinary information integrity research, especially new institutional infrastructure, educational pathways, and ethical frameworks.

The emerging field of information integrity research would benefit from greater coherence. Research and teaching in the area are often fragmented across disciplines, and groundbreaking research often emerges from interdisciplinary and multi-institutional collaborations. We recommend that the Working Group prioritize steps to build the field.

One promising direction is to invest in and facilitate the fundamental institutional infrastructure that is necessary for the field's success: new research centers, convenings, and publication venues that are specific to information integrity, reaching across disciplines and institutions. The Knight Foundation has emphasized this strategy in its recent grantmaking, and we encourage the Working Group to consider a similar model.

[14] Van Buren v. United States, 141 S. Ct. 1648 (2021).

Another important step would be developing educational pathways for students, including mid-career professionals, who may consider a career in information integrity research or practice. Because the field is still taking shape, educational offerings related to information integrity tend to be one-off courses rather than complete sequences of study. The federal R&D strategy could have a significant beneficial impact by supporting the development of new courses, curricula, degree programs, and academic concentrations that will build a pipeline of information integrity expertise.

Our final recommendation is that the Working Group address ethical considerations for the emerging area of information integrity research. Carrying out work in this field can involve collecting and analyzing personal information and can involve examining online platform practices without the platform's agreement. Information integrity research may also implicate speech protected by the First Amendment and may have a political valence. Ethical frameworks for information integrity research are essential, both to provide guidance for the conduct of research and to address foreseeable dilemmas before they occur. The PERVADE multi-institution collaboration, which has explored the ethics of social media research methods and is supported by NSF, is a potential model for how the federal strategy could address ethical considerations.

* * *

Thank you again for the opportunity to provide input to the federal government's strategic plan for information integrity research and development. We would be glad to provide additional detail or discussion as would be helpful to the Working Group.

Sincerely,[15]

Academic Researcher
Rutgers University

Academic Researcher
Center for Information Technology Policy, Princeton University

[15] We offer this comment as individual academic researchers.

Academic Researcher
Cornell University

Academic Researcher[16]
Princeton University

Academic Researcher
University of Washington

Academic Researcher
Stanford University

Academic Researcher
Princeton University

Academic Researcher
Northeastern University

[16] Principal author of this comment.

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

MITRE Corporation

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.

Response of The MITRE Corporation to the NITRD RFI on Federal Priorities for Information Integrity Research and Development

For additional information about this response, please contact:
Center for Data-Driven Policy
The MITRE Corporation

<<This page is intentionally blank.>>

About MITRE

MITRE is a not-for-profit company that works in the public interest to tackle difficult problems that challenge the safety, stability, security, and well-being of our nation. We operate multiple federally funded research and development centers (FFRDCs), participate in public-private partnerships across national security and civilian agency missions, and maintain an independent technology research program in areas such as artificial intelligence, intuitive data science, quantum information science, health informatics, policy and economic expertise, trustworthy autonomy, cyber threat sharing, and cyber resilience. MITRE's 8,000-plus employees work in the public interest to solve problems for a safer world, with scientific integrity being fundamental to our existence. We are prohibited from lobbying, do not develop or sell products, have no owners or shareholders, and do not compete with industry. Our multidisciplinary teams (including engineers, scientists, data analysts, organizational change specialists, policy professionals, and more) are thus free to dig into problems from all angles, with no political or commercial pressures to influence our decision-making, technical findings, or policy recommendations.

MITRE has provided unbiased, trusted advice to federal sponsors regarding how to understand and leverage information for decades, across applications both foreign and domestic. Over the past few years, we have also developed partnerships with research leaders in industry, academia, and the nonprofit sector to provide capabilities to prevent, detect, and respond to information integrity challenges. Our primary focus in this space has been on overcoming information integrity issues with likely dangerous outcomes, such as identifying the source or amplification of influence campaigns with divisive or violent intents or efforts to help states spot and overcome incorrect information about election processes and infrastructure.[1] We also have an interest in countermeasure development, especially in ways of increasing societal resilience through means such as inoculation against or prebunking[2] disinformation, counter-narrative development, and tuning social media platforms to create pro-social cascades.

Introduction

Individuals, organizations, and governments in open societies depend on access to trustworthy information to make good decisions for themselves, their companies, and their societies. The democratization of media and online content production brought about by the increased ubiquity of the internet has had many positive effects, but also has resulted in new challenges, including propagation of *misinformation* (false information), *disinformation* (false information intended to deceive), and *malinformation* (based on true information, but used out of context to mislead, harm, or manipulate)[3] – collectively referred to as mis/dis/malinformation in this response.

[1] MITRE SQUINT™ App Helps Election Officials in 11 States Spot Incorrect Election Information. 2020. MITRE, https://www.mitre.org/news/press-releases/mitre-squint-app-helps-officials-in-11-states-spot-incorrect-election-information. Last accessed April 29, 2022.

[2] D. Blackburn. "Policy Wrappers" for S&T Findings. 2022. MITRE, https://www.mitre.org/sites/default/files/publications/pr-22-1175-policy-wrappers-for-st-findings.pdf.

[3] Mis, Dis, Malinformation. 2022. Cybersecurity & Infrastructure Security Agency (CISA), https://www.cisa.gov/mdm. Last accessed April 26, 2022.

The CISA-based mis/dis/malinformation descriptors above include a concept of true versus false information, but in practice most information is usually not that clearly binary. While some items have been proved beyond a reasonable doubt (the day has 24 hours, the Earth is round and revolves around the Sun, etc.), many others have not. Science evolves, conditions change, the same object can look completely different from various perspectives, and results of assessments can vary wildly as the weighting of its various influences are minutely tweaked. Information integrity efforts must therefore not fall into the rabbit hole of trying to determine absolute truths but rather focus on better enabling entities to adequately assess information, determine its relevance, and properly leverage it via critical reasoning.

As public debate increasingly takes place on privately owned platforms, internet governance topics—including how best to understand, share, and react to online harms, appropriate transparency, and what role government might play in convening, coordinating, or providing oversight, regulation, or guidance to diverse stakeholders—form an important backdrop to the research and development needs for information integrity. Because of disagreement among experts on the drivers, impacts, and scale of information integrity problems, as well as the adjacency to politically charged discussions about free speech and data privacy, issues of information integrity must be addressed in an evidence-based, non-partisan, and scientific manner that is nonetheless attuned to the ongoing values-based discussion about what role communication technologies should play in our democratic society to ensure prosperity and security for all. In the spirit and intent of Integrated Deterrence, these measures and the capabilities that enabled them should also be considered in context of our nation's many allies and partners, who are also challenged with information integrity and are pursuing their own solutions.

We must also be cognizant of related issues that are detrimental to making progress on information integrity. For example:

- Some of the more vocal entities on the issue of mis/dis/malinformation are also prodigious developers or amplifiers of such information themselves.
- There are growing instances of advocacy or political activities that erroneously assess information as false when it doesn't align with their existing worldview or support their arguments.
- The concept of how to communicate about information integrity issues and activities so that they are accurately understood and embraced by the population remains in its infancy, with the establishment of the Department of Homeland Security's Disinformation Governance Board providing a recent high-profile example.

Questions Posed in the RFI

1. Understanding the information ecosystem: there are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

A common perspective for understanding and investigating the complex information ecosystem will be an important aspect of driving advancement of information integrity. Absent such a perspective, individual researchers may not fully recognize the various elements that should be considered while scoping their projects, nor the needs of those who will be impacted by the research. Key aspects that need to be included in such a common perspective include:

- Technical matters, such as infrastructure and data
- Motives and incentives
- Tactics, techniques, and procedures (TTPs)
- Attribution
- Spread/diffusion mechanisms (aka "information manuevers)
- Impact
- Mitigation
- Collaboration (cross-sector, interdisciplinary, and international)

The Department of Defense has taken some aligned initial actions when it added information as a seventh joint function, as the "information function encompasses the management and application of information and its deliberate integration with other joint functions to influence relevant-actor perceptions, behavior, action or inaction, and support human and automated decision making."[4] Additionally, the *DoD Concept for Operations in the Information Environment*[5] identifies the required capability to characterize and assess the informational, physical, and human aspects of the security environment. These concepts can be incorporated into a government-wide perspective. Additional research challenges in providing a common foundation for understanding information manipulation include:

Terminology. Common terminology is critical for any field's advancement as it enables every professional to represent, express, and communicate their findings in a manner that is effectively and accurately understood by their peers. While national and international vocabulary standards take time to develop, there is also a National Science & Technology Council (NSTC) precedent for establishing both definitions and directing agencies to consistently use and follow them on a priority science and technology topic.[6] The same approach could be leveraged here, starting with a collaborative research activity to develop the definitions.

Guidelines for Proper Scoping. Selecting a proper scope for a research project is critically important but can be challenging. While information ecologies (including people, practices, technologies, and values) are experienced locally, they can also have a broader reach that must be considered. For this reason, it is important that researchers define specifically the components of the information ecosystem that they intend to study, how, and to what ends. For example, studying social media issues via one social media platform, in one language, at a single point in time, while at times illuminating, requires contextualization within the larger set of information ecologies in which people interact—online and offline—across the globe. Given that information

[4] Doctrine for the Armed Forces of the United States. 2017. Department of Defense, https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp1_ch1.pdf.

[5] Joint Concept for Operations in the Information Environment. 2018. Department of Defense, https://www.jcs.mil/Portals/36/Documents/Doctrine/concepts/joint_concepts_jcoie.pdf

[6] In the mid-2000s, the NSTC's Subcommittee on Biometrics published a "Glossary" document of biometric terms. As part of its formal approval, its parent NSTC Committees also instructed member agencies to follow those definitions in their future activities. Non-governmental entities (mostly) aligned voluntarily as well and this document served as an important input in the development of an international vocabulary standard, which the subsequent *NSTC Policy for Enabling the Development, Adoption and Use of Biometric Standards* formally required agencies to follow.

ecosystems are vast and complex, it is incumbent on researchers to prudently plan, clearly define the scope and objectives of research, and then develop appropriate data collection and analysis methods, rather than simply beginning with a limited social media dataset out of convenience. Research that produces best practices and guidelines for properly scoping subsequent information integrity research would be helpful.

Data. Information integrity research requires data, which can be difficult to obtain and/or properly scope, and usually comes with personally identifiable information (PII) concerns. For example, social media researchers often use small Twitter datasets for convenience, but patterns identified in research may not accurately extend to other platforms. Differential access to platforms is also a key challenge, producing the effect of greater study on more developer-friendly platforms as opposed to those with the highest penetration and engagement. Creating adequate and appropriate datasets for training (and separately, testing) purposes, and can evolve over time, is needed. Policy on the use of publicly available information (or foreign websites) is also still needed to guide many federal departments and agencies.

Continuous Evolution. Longitudinal monitoring of the information ecosystem is challenging due to the complex nature of the environment (e.g., account attrition and platform migration, signal-to-noise issues), complicating model development and evolution. As a result, academic research on mis/dis/malinformation is often cross-sectional in nature, which limits applicability to use cases such as change detection (e.g., early identification of extremist behavior). Foundational work is needed to support monitoring that is longitudinal, multi-platform, and inclusive of a broad range of online and offline content.

Language and Platform Variance. One of the challenges in developing research that is multi-lingual, multi-platform, and longitudinal has been the wide variance in social media platform terms of service, policies, and practices. At present, there is a lack of transparency about information manipulation prevention, detection, and response efficacy and measurement within and across social media platforms. As a result, there is no cross-platform view of the information ecosystem as-is state for information manipulation research. There is also a lack of incentive for social media platforms to collaborate with external researchers and the federal government on information manipulation countermeasures research, or to increase transparency about the scope of information integrity issues on their platforms, the specific measures in place to prevent or counter those harms, and how those measures are performing over time. Research to overcome these gaps and to develop guidance on how to perform associated information integrity research from a common perspective is needed.

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. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?

Information manipulation is going to occur, both intentionally and unintentionally. It even shows up where we'd least expect it: somewhere between 10 and 20 percent of citations in peer-reviewed scientific journals are used to support claims that conflict with the findings of the original paper.[7] Thus, while efforts to minimize information manipulation can be helpful, efforts to support identifying and overcoming it will need to be a high priority.

In 2021, MITRE surveyed 29 recent mis/dis/malinformation research agendas and workshop proceedings from across academia, government, civil society, and industry, developing a meta-analysis of priority research needs. Themes within this meta-analysis, which summarize research priorities, are listed below. While this document has not yet been published, a pre-publication draft is provided as Appendix A of this RFI response for the NSTC's benefit.

- Data/infrastructure needs
- Attribution
- TTPs
- Motives
- Spread/diffusion
- Impact
- Mitigation
- Interdisciplinary/cross-sector collaboration
- Cross-platform
- Cross-format
- International perspectives

3. Information awareness and education: a key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

Manipulated information is pervasive because it works. The human mind is naturally receptive to information that aligns with one's cognitive biases. Social media makes it easy for everyone to further share this information with likeminded strangers around the world, which is also further promoted and leveraged by politicians, advocacy organizations, and traditional news media. Network theory shows us how information nodes can grow quickly, particularly with the support

[7] J. West and C. Bergstrom. Misinformation in and about science. 2021. Proceedings of the National Academy of Sciences of the United States of America, https://www.pnas.org/doi/10.1073/pnas.1912444117. Last accessed April 29, 2022.

of adversarial network manipulation. Technological solutions to these challenges are complex, while also raising free speech concerns. Overcoming manipulated information in the wild with counterarguments alone isn't a winnable approach due to realities highlighted in the Brandolini[8] and Finelli internet idioms.[9] The problem of manipulated information thus can't be solved without addressing the human component.

Two fundamental obstacles to addressing the human component of the manipulated information problem in modern populations are poor critical reasoning skills and a lack of desire to engage with those who hold alternative views,[10] both of which are routinely leveraged by advocacy-driven organizations. While research to overcome these obstacles won't directly solve the information integrity problem, it would provide insights on how to best raise the floor of what is conceivably possible and should thus be prioritized. We also need to recognize that both obstacles are problems that will require a long time to overcome, and thus sustained efforts will be required.

More directly focused lanes of needed research include:

- Cognitive bias and logical fallacy awareness and impact mitigation
- Intervention options and associated implementation considerations
- Novel education and communication approaches, including public-private partnerships, to inform community-level stakeholders
- How to reach citizens who have low critical reasoning skills and/or are unlikely to participate in digital literacy efforts
- How to properly measure the impact of information manipulation awareness efforts in both the short term and long term.
- Systematically defining and evaluating the conditions under which interventions of different kinds produce long-term positive outcomes
- Research to develop guidelines for critical thinking education by grade level in K-12

We should also pursue immediate efforts to help raise awareness and develop the information integrity skills of our population, both because of the urgent need but also to provide study matter for longer-term research. Example activities or other ideas to potentially leverage include:

- DARPA's Social Media in Strategic Communication (SMISC) Program[11] goal was to develop a new science of social networks built on an emerging technology base. Through the program, DARPA sought to develop tools to support the efforts of human operators to counter misinformation or deception campaigns with truthful information.

[8] Brandolini: The amount of energy needed to refute bullshit is an order of magnitude larger than to produce it.

[9] Finelli: An idiot can create more bullshit than an expert could ever hope to refute.

[10] The State of Critical Thinking 2020. Reboot, https://reboot-foundation.org/wp-content/uploads/_docs/Critical_Thinking_Survey_Report_2020.pdf.

[11] Social Media in Strategic Communication. 2022. DARPA, https://www.darpa.mil/program/social-media-in-strategic-communication. Last accessed May 10, 2022.

- DARPA's Influence Campaign Awareness and Sensemaking[12] program is developing techniques and tools that enable analysts to detect, characterize, and track geopolitical influence campaigns with quantified confidence.
- Integrate information integrity topics into existing cybersecurity competitions, such as the Air Force Association's CyberPatriot[13] program and the National Collegiate Cyber Defense Competition.[14]
- Identify existing integrity awareness offerings and make them accessible and easily findable to others, such as:
  - DoD's Influence Awareness course
  - Programs by USAID and CISA for local populations
  - Topic-specific offerings
- University of Washington course materials, and associated assets, which helps students identify manipulated data usage.[15]
- Create an ongoing public prize challenge to identify (and correct) each week's most noteworthy information manipulation.
- Create a website that highlights ridiculous information manipulations in an amusing manner, akin to an existing website[16] that aims to show that correlation is not the same as causation.

4. Barriers for research: information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

Research That Is Limited or in a Silo

Information integrity research and development requires a multidisciplinary approach. Bringing together a well-rounded team to research the intersection of technical, social, behavioral, and cognitive spaces is critical to making progress in this space. Much ongoing research, however, is performed within individual lanes of interest: technologists are largely interested in the

[12] B. Kellter. Influence Campaign Awareness and Sensemaking. 2022. DARPA, https://www.darpa.mil/program/influence-campaign-awareness-and-sensemaking. Last accessed May 10, 2022.

[13] Cyberpatriot – The National Youth Cyber Education Program. 2022. Air Force Association, https://www.uscyberpatriot.org/. Last accessed May 10, 2022.

[14] National Collegiate Cyber Defense Competition. 2022. Collegiate Cyber Defense Competition, https://www.nationalccdc.org/. Last accessed May 10, 2022.

[15] Syllabus with links to readings for course Calling Bullshit: Data Reasoning in a Digital World. 2022. University of Washington, https://www.callingbullshit.org/syllabus.html. Last accessed April 30, 2022.

[16] T. Vigen. Spurious Correlations. 2022. Tyler Vigen, http://www.tylervigen.com/spurious-correlations. Last accessed April 30, 2022.

information layer; societies are interested in the cognitive, social, and behavioral layer; and policy researchers are interested in broader implications.

- Relatedly, much of the blame for recent information integrity concerns has focused on social media; however, less than half of all Americans receive their news from this source[17]—and other sources, such as traditional media, are also substantial contributors to this national problem.

Remedy: Federal sponsorship of research activities (as well as related workshops or other events) should prioritize multidisciplinary investigations, and single-disciplinary investigations should still be done within the context of its placement within the developed common foundation for the information ecosystem. While implementing the NSTC's information integrity R&D strategy, they should similarly ensure interagency collaboration in each agency's project creation, implementation, assessment, and knowledge/capability transfer.[18,19]

Technical

Traditional information mediums—newspaper, radio, and television—operate on a technically static technology base. Internet mediums, however, rapidly evolve. Replicating studies in a technologically fluid environment is very challenging, but is also a key component of scientific evolution and trust in science.[20]

This fluidity also presents a data standardization challenge. Social media platforms, for example, have various ways of storing and representing data, as well as allowing access to that data, which complicates cross-platform research. These research challenges are further compounded because different populations use different platforms, with divergences driven both by purpose (socializing, shopping, or communicating) and subjects' age (younger individuals predominantly use different platforms than do older generations).

In the rush to understand how corrupt information spreads online, we believe baseline research related to this field was neglected. "Despite its prominence in the media, the study of disinformation is still in the process of answering definitional questions and hasn't begun to reckon with some basic epistemological issues."[21] This would explain why there are still disagreements about basic concepts and lexicon that are detrimental to a collaborative research environment.

Remedy: There needs to be serious attention devoted to standardizing (or translating) data and language across platforms and mediums within upcoming efforts to establish a common foundation for the information ecosystem. By creating structure on semantics and data, more progress may be made in the actual research space.

[17] M. Walker and K. Matsa. News Consumption across Social Media. 2021. Pew Research Center, https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/. Last accessed April 30, 2022.

[18] D. Blackburn. Interagency S&T Leadership. 2016. MITRE, https://www.mitre.org/sites/default/files/publications/pr-16-0916-interagency-s-and-t-leadership.pdf.

[19] J. Mervis. Spy Agencies Team Up with National Academies. 2016. Science, https://www.science.org/content/article/spy-agencies-team-national-academies. Last accessed May 1, 2022.

[20] D. Blackburn. When and How Should We "Trust the Science"?. 2021. MITRE, https://www.mitre.org/sites/default/files/publications/pr-21-1187-when-and-how-should-we-trust-the-science_0.pdf.

[21] J. Bernstein. Bad News: Selling the Story of Disinformation. 2021. Harper's Magazine, https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/. Last accessed May 1, 2022.

Social and Behavioral

In addition to the age-related platform use variances previously discussed, there's also no established way to evaluate the relationship between an individual's online activity and their real-world beliefs or behavior. Socio-cultural expertise is also limited, as R&D has predominantly focused on English-language materials. Also, the nation's polarization doesn't lend itself to large populations being willing to serve as study participants for information integrity research.

Remedy: Targeted research needs to be performed to overcome these barriers.

Adjacent Policies

Existing laws and policies that are themselves important (some barriers exist for a legitimate reason) can unfortunately create barriers for information integrity research, with data and privacy regulations being a prime example. While there have been attempts to leverage privacy-preserving methods to enable data sharing, needed characteristics within the datasets are often lost.

There is currently considerable debate and public awareness of potential first and fourth amendment concerns regarding federal government use of online materials. Recent policy announcements such as the Declaration for the Future of the Internet and DHS's creation of a Disinformation Governance Board can likewise generate concerns of government overreach in today's polarized environment, creating barriers for future federal research on the topic.

Remedy: Existing policies need to be thoroughly examined in the context of future information integrity research, with objectives to create guidelines and associated tools that allow permittable research to take place while simultaneously preserving the objectives of the original policies. Impacts of new policies need to be analyzed before they are established, and communications about these new policies must be both much better for general audiences as well as specifically tailored for this research audience.

5. Transition to practice: how can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

The Intelligence Community has demonstrated an ability to understand adversary information operations. Its iterative approach to implementing new concepts and capabilities could be instructive to broader public efforts and should be studied. Doing so would allow the private sector to sustain and build its disinformation programs, even as the development of new tools continues. The federal government can further help by brokering a public-private information sharing and R&D coordination mechanism to harness knowledge of the private sector while steering government-led research and operational integration.

6. Relevant activities: what other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

The meta-analysis provided in Appendix A includes references to a range of R&D strategies on this topic that will be useful here. Additionally, the community should study international efforts, results, and best practices to more fully understand the challenges of information integrity and help drive requirements for disinformation identification, diffusion, and attribution. As an example, The Institute for Strategic Dialogue,[22] an international organization headquartered in London, works to identify disinformation and other threats to democracy, to combat these threats, and to raise the awareness of governments to the challenge these threats represent. Understanding the views and activities of such organizations regarding the evolution of disinformation, its uses, and its effects can be a powerful adjunct to the development of the technologies our country needs to meet this challenge itself.

7. Support for technological advancement: How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission?

Since its inception (in the National Science and Technology Policy, Organization, and Priorities Act of 1976), OSTP has been an impactful force on federal—and, indeed, national—S&T activities. While each administration's OSTP has focused on different priorities, its two predominant functions have been consistent:

- **Policy for S&T:** Establish policies for federal S&T activities, help develop federal S&T budgets, and coordinate interagency S&T activities on priority topics.
- **S&T for Policy:** Ensure that S&T aspects are properly understood in other policy deliberations (such as in national security, government management, or economic matters).

Information integrity is an important aspect of both functions, not only in ensuring that OSTP's own activities and messages are proper but in helping overcome false assumptions by policymakers or inaccurate messages by external entities attempting to persuade them.

The federal information integrity R&D strategic plan will help catalyze research on information integrity matters, leading to new technological capabilities and social and behavioral science approaches to help non-experts better assess the messages they see. But these approaches usually are not rapidly or automatically placed into practice on their own. Rather, OSTP will need to actively work to ensure that they are through their own activities and influence. Doing so will not only help set examples for others to follow but also help ensure that OSTP is able to successfully meet both of its predominant functions with scientific integrity.

[22] Institute for Strategic Dialogue (ISD). 2022. ISD, https://www.isdglobal.org. Last accessed May 10, 2022.

Appendix A – Research Agenda Meta-Analysis

(PRE-PUBLICATION DRAFT)

Introduction and Methodology

We surveyed recent mis- and disinformation research agendas and priorities, as well as related conference and workshop proceedings from across academia, government, civil society, and industry. Our inclusion criteria for relevant materials were as follows: the publication or conference/workshop must be dated 2017 or more recent; the material must address mis- and disinformation research priorities (rather than specific policy or intervention recommendations); and the materials should preferably represent the interests of a group of researchers within a sector, rather than those of a sole lab. In total we reviewed 29 agendas and workshop summaries and identified researchers' key priorities for future work and included 22 here. A full annotated bibliography of included materials is available.

We identified several major themes from the literature which we present below. Direct quotes and citations from the literature are listed in each theme. Each bullet represents a stated research priority, need, or enabler.

Themes

Data/Infrastructure Needs

Infrastructure

- Bliss et al. point to the need for a common research infrastructure to obtain data from technology platforms, while preserving user privacy, following ethical guidelines, and protecting IP (Bliss, et al., 2020).
- A shared social, institutional, and technological infrastructure is necessary to develop datasets for studying the spread of misinformation on social media. This infrastructure can facilitate research and replicability but requires pressuring social media companies to share data (Lazer, et al., 2017).
- Research will continue to be hindered without broader access to historical data and to a wider range of platforms via APIs for Instagram, TikTok, WhatsApp, and YouTube (Pasquetto, et al., 2020).
- APIs are needed to access data relating to deletions, profile changes, 3rd – party application activities, abuse reports, and suspended accounts (Pasquetto, et al., 2020).
- API endpoints to show specific actions platforms take once a message is identified as containing misinformation (e.g., removals, warning labels, downranking) (Pasquetto, et al., 2020).

Data

- There is a need for both expression data (e.g., data corresponding to engagement with content such as likes and retweets) and impression data (e.g., data corresponding to people who read content). Impression data is hidden from researchers; access to this data would enable new research directions in studying both the spread of misinformation and the effectiveness of mitigation techniques (Pasquetto, et al., 2020).

- Demographic data about social media platform users can enable research into the spread, motivations for sharing, countermeasures against, and behavioral impacts of, misinformation (Pasquetto, et al., 2020).
- Fine-grained temporal data is required to characterize networks in which misinformation thrives, enable analysis of the types of events, policies, and technologies that are susceptible to mis/disinformation campaigns, and to aggregate characteristics about the populations that share misinformation (Pasquetto, et al., 2020).
- The research community would benefit from encrypted messaging data (e.g., WhatsApp data), specifically aggregated information about users and uses of the platform, viral and widely spread content, and random samples of names and groups (Pasquetto, et al., 2020).

Attribution

- Identifying who is conducting "social cybersecurity" attacks (Carley, 2020)
- Attribution is listed as one of the goals of detection of disinformation at scale by Bliss et al. (Bliss, et al., 2020)
- Understanding who shares misinformation would offer pathways to design and test interventions (Pasquetto, et al., 2020).
- Politicians, elites, and government officials are major (but understudied) sources of false information (Weeks & Zuniga, 2019).

TTPS

- Identify, track, and assess emerging propaganda and disinformation tactics, technologies, strategies, and patterns used by state and non-state actors to spread disinformation and propaganda across varied regions and cultures (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).
- Increased data access would enable improved characterization of misinformation in real-world contexts (Pasquetto, et al., 2020).
- Researchers are interested in accumulating the requisite data to study, detect, and combat manipulation. In regard to nation-state data, access to all of the organic Russian content from 2016 across various platforms can reveal TTPs and motives. For example, how much content was election related? What was troll behavior like in battleground states and across the nation? (Pasquetto, et al., 2020)
- Better understand strategies of international influence campaigns on target states (Goolsby & Montgomery, 2021).

Motives

- Understand what the perpetrators motive is; why the attack is being conducted (Carley, 2020).
- Understand why people share misinformation (e.g., does exposure to one's political opponents or allies affect willingness to share?) (Pasquetto, et al., 2020)
- Understand what misinformation is shared with whom and why (Pasquetto, et al., 2020).

- What motivates people to share information in particular when they know that the information is false? (Pasquetto, et al., 2020)
- Researchers can focus on answering underlying questions behind IOs: what are the motives behind a disinformation campaign? Why do people engage with problematic content? (V. Smith, 2020)

Spread/Diffusion

- Tracing and even predicting the spread of an influence campaign, including tracing attackers across multiple social media, monitors that suggest when diffusion is about to explode, peak, and peter out; improve theories of and methods for monitoring diffusion (Carley, 2020).
- Platforms can dampen the spread of information from just a few websites, the fake news problem might drop precipitously; steps by platforms to detect and respond to manipulations from bots and cyborgs will also naturally dampen the spread of fake news (Lazer, et al., 2017).
- Impression data is currently unavailable, but access to fine-grained impression data would allow researchers to measure the true reach of misinformation and could enable prediction of virality and diffusion path of misinformation with greater accuracy (Pasquetto, et al., 2020).
- Implement product design and policy changes on technology platforms to slow the spread of misinformation; researchers should prioritize understanding how people are exposed to misinformation (Murthy, 2021).
- Algorithms prioritize content that has or is expected to have, a high level of engagement. The risk is an overexposure of polarizing and controversial content and underexposure to less emotive but more informative content. Implications for policymaking include requiring online platforms to provide reports to users showing when, how, and which of their data is sold/used (Lewandowsky, et al., 2020).
- Exposure and belief in false information depends a lot on our social connections but we haven’t paid much attention to that yet. We don’t know the extent to which people are exposed to false information via their social connections or to what effect (Weeks & Zuniga, 2019).

Impact

Quantify the impact/influence/effect of misinformation

- Quantify by short- and long-term impacts through creation of improved measures of impact, such as polarization or mass-hysteria rather than traditional measures of reach such as number of followers, likes, and recommendation. (Carley, 2020)
- Precise, reliable, and validated measurement of the effect or impact of disinformation on communities- this requires formal statistical causal inference on human belief dynamics; requires advances in identification and extraction of complex cognitive/rhetorical structures and experimental sandbox representative of the ecosystem (Bliss, et al., 2020).
- Development of useful metrics of impact on a single and multiple platforms (Goolsby & Montgomery, 2021).

- Quantification of psychological impacts of disinformation/propaganda is a non-trivial challenge (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).
- Further research should be done that measures the impact of inauthentic behavior, better systems for recording these observations would enable this research. In addition, common metrics for measuring inauthentic behavior at scale should be developed (Wright, Stupak, Nikolich, Mattie, & J. Braun, 2020).
- Lack of common research standards are a concern: there is a lack of methods to measure the effects of influence operations (V. Smith, 2020).

Causal models

- Explore measurement, processes, and effects of polarization, particularly affective polarization (whether political, religious, ethnic, or another type). Particularly interested in causal models of polarization driven by informational, environmental, demographic, and institutional factors (Facebook Research, 2021).
- Expand research that deepens our understanding of health information and why it impacts people (Murthy, 2021).
- The research community has not well documented the effects of (or lack thereof) disinformation on outcomes we care about like voting, polarization, the rise of white nationalism, or echo chambers (Weeks & Zuniga, 2019).
- Politicians, elites, and gov officials are major sources of false information: regardless of whether people believe these claims, do members of the public respond to this information with more incivility, by becoming more engaged, or by ultimately becoming more polarized? (Weeks & Zuniga, 2019)

Offline/ “Real World” Influence

- Real-world influence efforts should be studied along with cyber-social efforts to better illuminate how real-world and cyber-world efforts converge, cohere, and amplify one another (Goolsby & Montgomery, 2021).
- Combine real-world study of human behavior with the study of cyber behavior in a diversity of local social contexts, investigating how social media engagement and participation in new social worlds result in the formation of different identities, beliefs, and behaviors that have significant implications for social stability within different systems of government (Goolsby & Montgomery, 2021).
- Consider both “hard influence” (influence that promotes the development of fissures in society) and “soft influence” (constructive, positive narratives and social rewards) (Goolsby & Montgomery, 2021)
- Measure and assess the impact and effect (including secondary effects) of propaganda and disinformation events on US and international audience decision making (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).
- There is a lack of rigorous research in the link between disinformation and resulting sub-optimal behaviors (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).

- Links between online engagement and radicalization remain unsubstantiated by data. Put more investment into understanding the problem, how it's manifesting, where it's manifesting, in order to be strategic (Brookie, Spitalnick, McCord, Rasmussen, & Gillum, 2021).
- Due to the difficulty of measuring 'real world' impact, much disinformation research has focused on spread using readily measurable behavior (clicks, retweets, site visits, etc.) (Colley, Granelli, & Althuis, 2020)
- Look beyond the spread of disinformation online and especially beyond social media, incorporate more diverse news-sharing behaviors, including offline. Focusing online on social media neglects the significance of traditional media and offline communication networks; social and traditional media should not be considered in isolation (Colley, Granelli, & Althuis, 2020).
- Research that explores deterrents to online and offline problematic behavior related to dangerous speech and harmful conflict (Facebook Research, 2021).

Impacts on Subpopulations

- Prioritize understanding how people are exposed to and affected by misinformation and how this may vary for different subpopulations (Murthy, 2021).
- Understanding how people navigate and trust information sources in specific contexts likely requires qualitative sociological and ethnographic research. The more community-specific research is, the better (Colley, Granelli, & Althuis, 2020).
- Understanding how people across different backgrounds, communities, and cultures interact with, are affected by, and decide to promote or share the spectrum of possibly problematic content (Facebook Research, 2021).

Mitigation

Countering and mitigating effects of disinformation through interventions

- Agent-based modeling to assess the relevant impact of interventions (Carley, 2020)
- External randomized control trials (RCT) on social media platforms without interference or involvement from the companies themselves would allow for rigorous research to understand what kinds of interventions are most effective at reducing an individual's propensity to share misinformation, and to what extent does revealing the source of factual interventions affect behavior (Pasquetto, et al., 2020).
- Research exploring the impact of interventions ideally involve randomly assigning users of platforms to various intervention versus control conditions. Specific interventions include labeling news headlines with fact-checking warnings, prompts that nudge users to consider accuracy before sharing, attempts to increase digital literacy, and assessing impact of incorporating layperson accuracy ratings (Pasquetto, et al., 2020).
- How do mitigation tactics such as removal, warning labels, and downranking, affect the way audiences respond to misinformation? (Pasquetto, et al., 2020)
- Which mitigation tactics work best with which audiences or demographics? (Pasquetto, et al., 2020)

- Development of new approaches to counter influence campaigns, including proactive and reactive strategies by U.S. and allies for messaging activities and other cyber-social efforts, as well as economic and other real-world approaches to counter influence (Goolsby & Montgomery, 2021).
- Study phases of disinformation operations to identify where technology-based solutions could be implemented (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).
- Passive fact checking may not be enough; rather we should think about corrective messages as a form of persuasion or social influence- we don’t yet know what message features are effective. Applying theories of social influence and persuasion can help by indicating what message elements are persuasive, which sources are credible, and how to reach less attentive audiences (Weeks & Zuniga, 2019).

Building community resiliency to attacks

- Scalable techniques for teaching critical thinking for social media (Carley, 2020).
- Basic research on the characteristics of resilient communities (Carley, 2020).
- What are the necessary ingredients for social information systems to encourage a culture that values and promotes truth? (Lazer, et al., 2017) Educational efforts, sourcing debunking from communities who maintain a shared narrative, use of social pressure, form bridges across communities to foster production of more neutral and factual content, understanding that not all individuals will be susceptible to intervention.
- Need to understand ways in which a common ground for evidence and rules of arguments can be re-established (Pasquetto, et al., 2020).
- Explore the relation between digital literacy and vulnerability to misinformation; including studies of individuals, small groups, and larger communities, but also wider inquiries into factors that shape the context for the user experience online (Facebook Research, 2021).
- Programs to develop digital literacy to identify disinformation in the wild (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021).
- Equip Americans with the tools to identify misinformation, make informed choices about what information they share, and address health misinformation in their communities in partnership with trusted local leaders (Murthy, 2021).
- Invest in longer-term efforts to build resilience against health misinformation, such as media, science, digital, data, and health literacy programs and training for health practitioners, journalists, librarians, and others. (Murthy, 2021)
- Strengthen and scale the use of evidence-based educational programs that build resilience to misinformation (Murthy, 2021)
- Establish quality metrics to assess progress in information literacy. (Murthy, 2021)
- Tools for digital literacy, particularly among older generations, may be useful in limiting the spread of inauthentic behavior. (Wright, Stupak, Nikolich, Mattie, & J. Braun, 2020)

- Educate the public on trustworthy digital information: establish a grant program led by the NSF for the purpose of developing a curriculum on trustworthiness of information in the digital age. (Goodman, Carlson, & Bray, 2021)
- The effectiveness of fact-checking and social interventions is questionable; going forward the field aims to identify social factors that sustain a culture of truth and design interventions that help reward well-sourced news. (Lazer, et al., 2017)

Strengthening institutions

- Support efforts to strengthen local reporting in the face of tightening budgets through subsidies for local news outlets and help obtaining non-profit status. (Lazer, et al., 2017)
- Help people understand the rigor that goes into journalism (source gathering, fact checking, no surprises policy). (Goodman, Carlson, & Bray, 2021)
- Government has a responsibility to encourage independent, professional journalism, avoid crackdowns on the news media's ability to cover the news, avoid censoring content and making online platforms liable for misinformation. (West, 2017)
- Understanding how people navigate and trust information sources in specific contexts is critical, and likely requires offline ethnographic research to address community-specific concerns. (Colley, Granelli, & Althuis, 2020)

Interdisciplinary/Cross-Sector Collaboration

- Outreach and collaboration across academic institutions is needed. Currently this is being facilitated in large part by the DoD Minerva program and the Knight foundation (Carley, 2020).
- The next generation of technologists need to be trained in applied ethics so that their processes align with a practicable mindset and toolset (Bliss, et al., 2020).
- Collaboration between conservatives and liberals to identify bases for factual agreement will heighten the credibility of counter-misinformation endeavors (Lazer, et al., 2017).
- Find ways to support and partner with the media to increase the reach of high-quality, factual information.
- Scholarship that proceeds without acknowledging the theoretical framework of propaganda, analysis of ideology and culture, notions of conspiracy theory, and concepts of misinformation and impact, does so to its empirical detriment and makes identifying solutions harder to articulate because the actual problem to be solved is unclear (Anderson, 2021).
- Kenneth Joseph, Nir Grinberg, and John Wihbey have identified broad avenues of future collaboration between technology platforms and the academic community, ranging from frameworks for survey studies, user-level data about misinformation interactions, actions taken by platforms, content curation and moderation algorithms, and access to historical data (Pasquetto, et al., 2020).
- Miriam Metzger and Andrew Flanagin likewise identified areas of collaboration between social media platforms and academic researchers to capture user and network level data.

- Albarracin et al point out that interdisciplinary and intersectoral collaborations between government and social media companies is necessary, particularly when social media companies are not always forthcoming about the information they are spreading.
- Both government and industry have called for social-science-forward approaches to drive the next generation of computational research into quantification of disinformation and its community impacts.
- Facebook has called for collaboration within academic disciplines, with particular attention paid to social science methods, comparative politics and cultural research, and studies that focus on Non-Western measures and analyses (Facebook Research, 2021).
- Topic 4 of the Minerva 2021 research priorities seeks multidisciplinary theoretically innovative approaches from disciplines such as anthropology, cross-cultural sociology, political science, political economy, and cross-cultural social psychology, working in collaboration with computer and information sciences to develop a social-science-forward approach to the development of social theory and the creation of new techniques needed to carry out a systemic analysis of social influence in online and offline cross-cultural milieus, cyber-social dynamics, narrative, and in languages other than English. Particular attention is paid to studies in important strategic regions in Asia, Africa, and Latin America. (Goolsby & Montgomery, 2021)
- Topic 5 of the Minerva 2021 research priorities seeks to involve social scientists, media researchers, area specialists working with information and/or scientists to develop approaches to studying influence of both online and offline communities. (Goolsby & Montgomery, 2021)
- Topic 6 of the Minerva 2021 research calls for new models of collaboration, innovative experimental design, and data analysis to explore computational social science research on difficult-to-access environments. (Goolsby & Montgomery, 2021)
- Global Engagement Center Counter Disinformation and Propaganda center identified primary shortfalls in two areas: understanding the information environment and measuring impact and effectiveness, suggesting that basic capabilities to develop an awareness of propaganda and disinformation may be low, and that there is limited interagency and intergovernmental capability coordination. (Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center, 2021)
- In the U.S. Surgeon General’s Advisory on Confronting Health Misinformation (Murthy, 2021), one key area of concern was to convene federal, state, local, territorial, tribal, private, nonprofit, and research partners to explore the impact of health misinformation, identify best practices to prevent and address it, issue recommendations, and find common ground on difficult questions, including appropriate legal and regulatory measures that address health misinformation while preserving user privacy and freedom of expression.
- Funders and foundations are encouraged to move with urgency toward coordinated, at-scale investment to tackle misinformation, and to incentivize coordination across grantees to maximize reach, avoid duplication, and bring together a diversity of experience.

- Government agencies are encouraged to convene federal, state, local, territorial, tribal, private, nonprofit, and research partners.
- Projects and development should be informed by anthropologists, with particular attention paid to differences in ways individuals interact with information. (Wright, Stupak, Nikolich, Mattie, & J. Braun, 2020)

Cross Platform

- Inspect information practices and flows across multiple communication technologies or mediums. In particular, individual, group, and community effects of information campaigns, inauthentic behavior, or coordinated activities across multiple communities, networks, channels, or platforms. (Facebook Research, 2021)
- Platforms should provide vetted researchers with comparable, open APIs to enable cross-platform analytics. (Bliss, et al., 2020)
- Are some types of events, policies, or technologies uniquely susceptible to mis/disinformation campaigns? (Pasquetto, et al., 2020)
- Can we enable research using private encrypted messages, such as those on WhatsApp to track the dissemination of misinformation at scale? (Pasquetto, et al., 2020)
- Cross-platform information ecosystem understanding: research that inspects information practices and flows across multiple communication technologies or mediums. (Facebook Research, 2021)
- Researchers need to look to non-obvious platforms (e.g., Airbnb). (Brookie, Spitalnick, McCord, Rasmussen, & Gillum, 2021)

Cross-Format

- Investigate the role of non-textual media (images, videos, audio, etc.) on the effectiveness of and people’s engagement with misinformation. Media types include basic multimedia, simple or advance manipulated multimedia (deepfakes, cheapfakes), out-of-context imagery, impersonation of public figures/organizations, etc. (Facebook Research, 2021)

International Perspectives

- Comparative research and inclusion of non-Western regions that have experience a growth in social media platform use: particularly South and Central America, Sub-Saharan and Northern Africa, the Middle East, and Central, South and Southeast Asia. (Facebook Research, 2021)
- Utilize non-Western measures and analyses to study affective polarization, particularly when applied to questions of equitable impact on vulnerable communities. (Facebook Research, 2021)
- Analyze dangerous speech, conflict, and violence in markets with limited institutions, developing media markets, and various levels of democracy in non-Western contexts.
- Research has lagged in studying important strategic regions in Asia, Africa, and Latin America. Multidisciplinary theoretically innovative approaches are needed to carry out systemic analysis of social influence in online and offline cross-cultural milieus and in languages other than English. (Goolsby & Montgomery, 2021)

- Enhance understanding of difficult-to-access environments, such as those experiencing enduring conflicts to societies that broadly restrict researcher access. (Goolsby & Montgomery, 2021)
- Overseas disinformation trends incubate in other countries before coming to the U.S.; likewise, U.S. disinformation exports to other countries (e.g., anti-vaxx content). (Frenkel & Newton, 2021)
- Other countries could use the U.S.'s seizure of Iranian controlled domains as an opportunity to use claims of disinformation to silence dissent. (Frenkel & Newton, 2021)
- The global, long reach of extremism leads to groups finding new issue of the day. (Brookie, Spitalnick, McCord, Rasmussen, & Gillum, 2021)
- Not all disinformation campaigns are aimed at causing violence; some seek to rewrite history. (Brookie, Spitalnick, McCord, Rasmussen, & Gillum, 2021)
- Researchers need to look also at non-Western dominant brand names (e.g., WeChat). (Brookie, Spitalnick, McCord, Rasmussen, & Gillum, 2021)
- Little work is done to understand how governments use influence operations on their own citizens, the role of media in these campaigns, influence in Africa, and non-English language influence operations. (V. Smith, 2020)
- In many countries, few individuals share news on social media, fewer trust it, and the trust is declining. (Colley, Granelli, & Althuis, 2020)
- Study disinformation's impact in a broader range of cultural contexts. (Colley, Granelli, & Althuis, 2020)
- There are 83+ languages in Ethiopia; as of June there was only automated translation available for Amharic (the national language) and neither human or automated translation for any other languages. The platforms could not see/hear/understand the content circulating on their platforms. (Ingrahm, 2021)

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review, 12*(6), 1043-1047.

Anderson, C. (2021, April 15). Propaganda, misinformation, and histories of media techniques. *Harvard Kennedy School (HKS) Misinformation Review.* doi:https://doi.org/10.37016/mr-2020-64

Atlantic Council GeoTech Center & U.S. State Department Global Engagement Center. (2021). Second Counter Propaganda and Disinformation Technology Working Group. unpublished.

Bliss, N., Bradley, E., Garland, J., Menczer, F., Ruston, S. W., Starbird, K., & Wiggins, C. (2020). *An Agenda for Disinformation Research.* Retrieved from arXiv preprint: http://arxiv.org/abs/2012.08572

Brookie, G., Spitalnick, A., McCord, M., Rasmussen, N., & Gillum, R. (2021, June 22). Fighting online extremism in 'the Klan den of the twenty-first century. Atlantic Council DFR Lab 360/Open Summit. Retrieved from https://www.atlanticcouncil.org/blogs/new-atlanticist/how-to-spot-the-latest-trends-in-digital-disinformation/

Carley, K. M. (2020). Social cybersecurity: an emerging science. *Comput Math Organ Theory*, *26*, pp. 365-381. doi:https://doi.org/10.1007/s10588-020-09322-9

Colley, T., Granelli, F., & Althuis, J. (2020). Disinformation's Societal Impact: Britain, COVID, and Beyond. *Defence Strategic Communications, 8*. doi:10.30966/2018.RIGA.8

Facebook Research. (2021). *Foundational Integrity Research: 2021 Misinformation and Polarization request for proposals.* Menlo Park: Facebook. Retrieved August 11, 2021, from https://research.fb.com/programs/research-awards/proposals/foundational-integrity-research-2021-misinformation-and-polarization-request-for-proposals/

Facebook Research. (2021, June 14). Foundational Integrity Research: 2021 Misinformation and Polarization request for proposals. Retrieved August 11, 2021, from https://research.fb.com/programs/research-awards/proposals/foundational-integrity-research-2021-misinformation-and-polarization-request-for-proposals/

Feinman, S., & Entwisle, D. R. (1976). Children's ability to recognize other children's faces. *Child Development, 47*(2), 506-510.

Fredheim, R., Bay, S., Dek, A., Dek, I., & Singularex. (2021). Social Media Manipulation Report 2020. NATO Strategic Communications Center of Excellence. Retrieved from ISBN: 978-9934-564-92-5

Frenkel, S., & Newton, C. (2021, June 22). How to spot the latest trends in disinformation. Atlantic Council DFR Lab 360/Open Summit. Retrieved from https://www.atlanticcouncil.org/blogs/new-atlanticist/how-to-spot-the-latest-trends-in-digital-disinformation/

Goodman, J., Carlson, T., & Bray, D. (2021). *Report of the Commission on the Geopolitical Impacts of New Technologies and Data.* The Atlantic Council GeoTech Center. Retrieved from ISBN-13: 978-1-61977-178-9

Goolsby, R., & Montgomery, D. (2021). *FY21 Minerva Research Initiative Topics of Interest.* Minerva Research Initiative. Retrieved from https://minerva.defense.gov/Owl-In-the-Olive-Tree/Owl_View/Article/2069926/it-takes-social-science-to-counter-the-power-of-russias-malign-influence-campai/

Grother, P. J., Ngan, M. L., & Hanaoka, K. K. (2018). *Ongoing face recognition vendor test (FRVT) part 2: Identification.* Gaithersburg, MD, USA: NIST.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects.* Gaithersburg, MD, USA: National Institute of Standards and Technology.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Ongoing face recognition vendor test (FRVT) part 1: Verification.* Gaithersburg, MD, USA: NIST.

Ingrahm, M. (2021). What should we do about the algorithmic amplification of disinformation. *Columbia Journalism Review*. Retrieved from https://www.cjr.org/the_media_today/what-should-we-do-about-the-algorithmic-amplification-of-disinformation.php

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating Fake News: An Agenda for Research and Action. *Combating Fake News: An Agneda for Research and Action.* Cambridge. Retrieved from https://www.sipotra.it/wp-content/uploads/2017/06/Combating-Fake-News.pdf

Lewandowsky, S., Smilie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., . . . Leiser, M. (2020). *Technology and Democracy: Understanindg the influence of online technologies on political behavior and decision-making.* European Commission Joint Research Centre (JRC). doi:http://dx.doi.org/10.2760/593478

Murthy, V. H. (2021, July 15). Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment. Retrieved August 12, 2021, from https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf

Nothhaft, H., Bradshaw, S., & Neudert, L. (2018). Government Responses to Malicious Use of Social Media. Riga: NATO Strategic Communications Center of Excellence. Retrieved from ISBN: 978-9934-564-31-4

Pasquetto, I. V., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., . . . Flammini, A. (2020, December 9). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*. Retrieved August 11, 2021, from https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139.

V. Smith, N. T. (2020). *Survey on Countering Influence Operations Highlights Steep Challenges, Great Opportunities.* Carnegie Endowment for International Peace Partnership for Countering Influence Operations. Retrieved from ttps://carnegieendowment.org/2020/12/07/survey-on-countering-influence-operations-highlights-steep-challenges-great-opportunities-pub-83370

Weeks, B., & Zuniga, H. G. (2019). What's Next? Six Observations for the Future of Political Misinformation Research. *American Behavioral Scientist.* doi:https://doi.org/10.1177/0002764219878236

West, D. (2017). *How to combat fake news and disinformation.* Brookings. Retrieved from https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/

Wright, A., Stupak, P., Nikolich, A., Mattie, K., & J. Braun. (2020). Inauthentic Behavior in Online & Digital Systems Final Conference Report. *Inauthentic Behavior in Online & Digital Systems.* University of Chicago. Retrieved from https://harris.uchicago.edu/files/nsf_report_final.pdf

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica, 114*(1), 101-114.

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

Mozilla

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.

moz://a

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

RESPONSE OF MOZILLA TO REQUEST FOR INFORMATION ON FEDERAL PRIORITIES FOR INFORMATION INTEGRITY RESEARCH & DEVELOPMENT

Table of Contents

1. About Mozilla

Mozilla is a unique public benefit organization and open source community formed as a non-profit foundation in the United States. We have a strong reputation for our commitment to ensuring that privacy and security are fundamental to the internet. This is one of our guiding principles that recognises, among other things, that the internet is integral to modern life; the internet must remain open and accessible; security and privacy are fundamental; and that a balance between commercial profit and public benefit is critical.[1] These principles, in addition to our Data Privacy Principles[2], provide the basis for the way we develop products, manage the consumer data we collect, how we select and interact with partners, and how we shape our public policy and advocacy work.

Our Public Mission & Incentives

Mozilla's story originated in 1997 with Netscape Navigator, the original consumer browser and a popular browser of the 1990s. In a historic move for competition, Netscape publicly released its new browser engine (called "Gecko") under an open source license to enable others to verify, improve, and reuse the source code in their own products. The company was later the subject of the failed acquisition strategy of a powerful digital gatekeeper, when AOL purchased it in 1999. Although Netscape did not last following its acquisition by AOL, its open source browser engine Gecko has continued to shape the internet.

The non-profit Mozilla Foundation was created in 2003 to continue work on open source browser technology and with a larger mission to preserve the open internet. Firefox v1.0 was released in 2004 using Gecko with volunteer open source code contributions from around the world, and it was one of the first major consumer facing products to be built in this way using open source methodology. Today localization developers continue to make Firefox available in local languages and with local customizations for their communities to access the internet. Other developers have forked the Firefox codebase and used the Gecko browser engine to create new browsers with different features. The most well known example is Tor, an anonymity browser frequently used by journalists and human rights activists. While it has officially been blocked in Russia,[3] reliance on Tor has increased recently as a means to gain access to the open internet.[4]

In 2005, the Mozilla Foundation created a wholly-owned taxable subsidiary, the Mozilla Corporation, to serve its public mission through open source technology and product

[1] Mozilla’s 10 Principles, https://www.mozilla.org/about/manifesto/
[2] Mozilla’s Data Privacy Principles, https://www.mozilla.org/en-US/privacy/principles/
[3] Maria Xynou, Arturo Filastò. Russia Started Blocking Tor. OONI, December 17, 2021. https://ooni.org/post/2021-russia-blocks-tor/
[4] Sam Schechner and Keach Hagey. Russia Rolls Down Internet Iron Curtain, but Gaps Remain. WSJ, March 12, 2022. https://www.wsj.com/articles/russia-rolls-down-internet-iron-curtain-but-gaps-remain-11647087321

development of Firefox. In addition to remaining the sole shareholder of the Corporation, the Foundation advocates for better privacy, trustworthy AI, and digital rights and runs philanthropic programs in support of a more inclusive internet. These programs currently include fellowships and awards that invest in community leaders who are developing technology, policy, education and norms that will ultimately protect and empower people online.

2. Greater Transparency into Hidden Harms

A great amount of harm, including but not limited to the effects of disinformation, happens on major tech platforms outside the view of regulators and the public. These platforms offer highly sophisticated targeting tools that allow peddlers of disinformation to narrowly segment their audience, tailor content accordingly, and reach people most susceptible to their messages. Each person has their own individualized, potentially misleading experience.

This highly personalized experience means that harm enabled by platforms through their targeting systems is not easily identified by regulators, watchdog groups, or researchers. Because the experience is so personalized, harm can only be shown anecdotally, rather than systematically. There is dangerously little insight into what people experience, what ads are presented to them and why, and what content is recommended and why. This creates an asymmetry of information between those who produce disinformation and those seeking to understand it.

To address this, we need greater access to platform data (subject to strong user privacy protections), greater research tooling, and greater protections for researchers. This is why Mozilla has invested in building tools for researchers and why we support legislative solutions to provide greater insight into online disinformation, discrimination, and deception currently hidden from the public and from regulators.

The remainder of this submission discusses these gaps in more detail and is responsive to specific questions in the Request for Information concerning key research challenges, barriers for conducting information integrity R&D, and support for technological advancement in the fields of measurement and research platform development.

3. Robust Research Platforms are Essential to Understanding the Information Ecosystem & Protecting Information Integrity

The ability to understand public life on the internet is largely concentrated in the hands of private actors. These for-profit companies have a vested interest in perceptions of public life online. Such interests raise questions from external researchers, among many others, about disinformation on platforms and whether such platforms fairly disclose information that ensures better accountability.

Unfortunately, the data and tools made available by major platforms to understand topics like threats to information integrity remain inadequate. Most of the voluntary public-facing measures by major platforms have failed. Researchers have found, for example, that ad transparency tools are often nearly unusable.[5] At the same time, major companies continue to threaten legal action against good faith security research into disinformation.

We need far more robust research tools along with a deep pool of subject matter experts capable of taking advantage of these tools. More specifically, the software projects and infrastructure necessary to understand online public life require significant domain expertise, technical expertise, and capital to build and maintain. Efforts like the Markup's CitizenBrowser project, or New York University's Ad Observer have noted their substantial startup costs, as well as ongoing operational costs. These barriers to entry consequently mean that the capability to study online public life remains accessible to few organizations.

The research produced by these teams has begun to shift the understanding and perceptions of legislators, regulators, and the public. New regulations across the globe, such as the General Data Protection Regulation (GDPR)[6], California Consumer Privacy Act (CCPA)[7], and the Digital Services Act (DSA)[8], are starting to move the broader data economy to center on informed user consent for data collection, and user control of the data collected from them.

Mozilla sees an opportunity to lower barriers for researchers to study and understand life online, and to provide users and the public with consent-driven approaches to exchanging their data.

Mozilla's Rally Project - What it is and How it Works

To explore these opportunities, Mozilla launched Rally, an opt-in platform for consumers to donate their data to researchers and causes they support. Mozilla envisions Rally as a platform through which users can affirmatively control their data and how it's used.

In its pilot phase, Rally has worked with journalists, academics, and non-profit researchers to develop and launch projects to understand the public's experiences online through user contributed browser and interaction data. The browser is a critical tool to allow people to navigate the online world. It can therefore provide significant insight into what people are experiencing online, what information they are consuming,

[5] Laura Edelson. Facebook's political ad spending numbers don't add up. Medium, October 12, 2020. https://medium.com/online-political-transparency-project/transparency-theater-facebooks-political-ad-spending-numbers-don-t-add-up-d7a85479a002
[6] The EU General Data Protection Regulation, https://gdpr.eu/
[7] The California Consumer Privacy Act, https://oag.ca.gov/privacy/ccpa
[8] The EU Digital Services Act, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

and how they are being manipulated, as long as that data can be collected in a privacy respecting manner. Rally aims to do just that, taking advantage of instrumentation tools already in the browser to potentially power research and studies by reputable parties outside Mozilla.

Current Rally studies[9] include one focused on “Political and COVID-19 News” across the web. This study can help us understand how web users are exposed to, consume, and share these types of information, which can inform efforts to distinguish trustworthy and untrustworthy content. Another study focuses on local news. The results will help build our understanding of how the modern news environment works, and which alternative funding models for local journalism may be feasible. These studies demonstrate the potential power of research platforms like Rally to contribute to information integrity R&D.

Users can sign up for Rally, select the projects they would like to contribute data to, and are prompted to install a browser extension in order to participate. Projects launched on Rally are reviewed by the Mozilla Rally team in order to vet the design and implementation of the studies so that users can have the confidence that Rally studies adhere to Mozilla’s data practices.

The web extensions built for Rally projects rely on Rally’s software development toolkits to ensure that data measured on a user’s machine is encrypted and transmitted securely to the Mozilla data platform. From there, data is placed in each project database, where only a restricted list of researchers and Mozilla Rally staff are permitted access. Researchers then access and run analyses on row level user data on the Mozilla data platform.

4. Vetting Qualified Researchers & Proposals

For research platforms like Rally, granting access to the *right* parties is important to ensuring our tools are not abused and users are not placed at risk. Currently the burden is on the Rally team to determine what researchers to work with, what credentials researchers need in order to gain access to our platform, and what specific research proposals are in the public interest. This is a responsibility we take seriously. To make this determination, we review both the researchers and their proposed studies. Unfortunately, this level of diligence can be cumbersome, limiting our ability to offer Rally to a large number of researchers.

This diligence burden currently must be independently borne by each research platform seeking to support work on information integrity. Further, research vetting may present prohibitive barriers to other companies or other research platforms that lack Mozilla’s

[9] Mozilla Rally current studies, https://rally.mozilla.org/current-studies/

technical expertise, resources, and mission focus, making them less willing to take on this burden, less likely to offer research tools, and more likely to make mistakes when they do so.

Unlocking the potential of solutions like Rally and allowing them to offer tooling at scale to a diversity of researchers requires relieving this burden on individual parties and companies. These tools are potentially a shared resource and the diligence burden ideally could be shared rather than duplicated. To address this need, there should be an independent or governmental body that can take on responsibility for this vetting. Current proposals in Congress[10] envision just such an approach, asking the National Science Foundation (NSF) to establish a process to solicit research applications and vet qualified researchers. We encourage NSF to explore creating such a program, even absent a legislative mandate. At a minimum, the Federal government through the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) and the NSF could work to create consensus standards that companies like Mozilla adopt and use.

Such an adjudicative body or standards need to address the following two areas:

- **Researcher Vetting.** Currently Mozilla's Rally platform is available to a very small number of researchers, mostly at marquee universities, allowing us to use university reputation as a proxy for researcher credibility. While this approach helps manage risk effectively, it has obvious shortcomings and limits the number of researchers that may benefit from Rally. We encourage the NSF and NITRD NCO to explore how to establish some type of standard or credentialing process, one not strictly limited by university affiliation, to vet researchers prior to their application to the research platform. Such a credential would be an incredibly valuable signal that Mozilla and others could use as criteria to determine who gains access.

- **Public Interest Value, Study Ethics & Methods**. Separate from researcher vetting, there must be standards established for research proposals themselves to ensure they are ethically designed and would contribute to the public interest. We recognize that, despite Mozilla's long track record of work to build a healthier Internet, we should not be the sole arbiter of what constitutes public interest research. Nor are we necessarily the right party to judge the ethics of the studies or the soundness of their scientific methods. This is a function better served by mechanisms like Institutional Review Boards (IRBs), which are today available to university affiliated researchers. There should be an independent evaluation

[10] Platform Accountability and Transparency Act (PATA), https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf

process that can assess both the soundness of the method and the public interest value of the research.

5. Policy Tools to Address Threats to Information Integrity

Robust research platforms are not, however, enough. Legislative or regulatory action is needed to complement and support independent research into disinformation. Recent history has shown that major tech platforms do not have sufficient incentive to provide the necessary level of transparency and access to researchers and that they must be required to do so.[11] To that end, Mozilla's policy advocacy focuses on two particular areas: researcher safe harbor and ad transparency. We strongly support provisions of the European Union's Digital Services Act[12] that would mandate disclosure of all ads on tech platforms. We are similarly encouraged by recent Congressional proposals that seek to provide accountability and transparency into the real world impact of disinformation, such as the Platform Accountability and Transparency Act (PATA)[13], which would provide important protections for researchers and require disclosure of ads.

Mandating a Safe Harbor to Protect Public Interest Researchers

Public interest research is key to shedding light on hidden harms and disinformation. Experts engaged in this research need to be protected. Accordingly, Mozilla has called for a safe harbor to allow researchers, journalists, and others to study disinformation and access relevant datasets, free from threat of legal action.

Mozilla often hears of researchers who are concerned that companies or governments may take legal action against them for their legitimate research – including civil or criminal penalties under laws such as the Computer Fraud and Abuse Act (CFAA), violations of Terms of Service, and more. Facebook, for example, has blocked research tools and threatened legal action against researchers seeking to investigate election integrity and misinformation online.[14]

These actions by platforms not only put researchers themselves at legal risk, but also stifle vital transparency into real world harm by deterring individuals and institutions doing critical work. Indeed, research tools and initiatives most vital to public interest -

[11] Marshall Erwin. Why Facebook's claims about the Ad Observer are wrong. Mozilla blog, August 2021. https://blog.mozilla.org/en/mozilla/news/why-facebooks-claims-about-the-ad-observer-are-wrong/

[12] Owen Bennett. Mozilla publishes position paper on EU Digital Services Act. Mozilla Open Policy & Advocacy blog, May 18, 2021.
https://blog.mozilla.org/netpolicy/2021/05/18/mozilla-publishes-position-paper-on-eu-digital-services-act/

[13] PATA, https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf

[14] Jeff Horwitz. Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting. WSJ, October 23, 2020.
https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533

most capable of identifying patterns of harm or threats to information integrity on major tech platforms - may receive the greatest scrutiny and be subject to the greatest legal exposure. This is the pattern we have seen with New York University's Ad Observatory project, which has offered research tools effective at identifying harms on tech platforms, and as a result, has had to withstand sustained legal attack.

The risk of legal exposure, both for ourselves and our research partners, is something that Mozilla must be mindful of when offering a research platform like Rally. In Mozilla's case, we have the legal expertise and resources such that we cannot be intimidated by spurious legal threats from major platforms. This is unlikely to be the case, however, for many of our potential research partners.

A safe harbor would protect and promote research in the public interest as long as researchers handle data responsibly and adhere to professional and ethical standards, such as those developed to support the vetting process described above. There is enormous value this can provide to the public. Mozilla has one of the earliest Bug Bounty programs in software. We make clear that we will not threaten or bring any legal action against anyone who makes a good faith effort to comply with our vulnerability notification policy because this encourages security researchers to investigate and disclose security issues. Their research helps make the internet a safer place.

Ad Transparency to Combat Hidden Harm & Disinformation

Advertisements are a critical vehicle for dissemination of disinformation. Research has found ad targeting helps peddlers of disinformation find and build their initial audience in the early stages of a disinformation campaign. Once that audience has been created, organic growth then takes over.

The public and regulators need far greater access to data about ads and ad targeting. To address this gap, Mozilla supports Executive or Congressional action requiring social media advertisers and companies to publicly disclose the ads and targeting criteria appearing on social media platforms. Ad disclosure should include the following:

- Apply to all advertising, so as not to be constrained by arbitrary boundary definitions of 'political' or 'issue-based' advertising;
- Include disclosure obligations that concern advertisers' targeting parameters for protected classes as well as aggregate audience demographics, where this makes sense given privacy and other considerations;
- Provide Broad Access to Data. Data should be broadly available to regulators, researchers, journalists, and watchdog groups, rather than restricting access to privileged stakeholders.

Previous regulatory initiatives aiming at ad transparency to combat disinformation have generally focused on ‘political’ advertising. Focusing on purely ‘political’ advertising (e.g. advertising copy developed by political parties) is too narrow and is insufficient to capture the complex web of actors involved in politically-motivated disinformation online. Moreover, a broad ads disclosure framework could also drive transparency with respect to what is known as ‘issue’ advertising. Experience has shown how disclosure obligations that include this broader category of political ads put platforms in a challenging position, requiring them to decide what is 'political' in nature, which can vary depending on context, jurisdiction, and time. A focus on all ads would negate this line-drawing challenge.

Further, the inclusion of all ads allows for the identification and analysis of other forms of systemic harm that may be occurring in the current ad ecosystem. Indeed, other types of advertising that are not overtly political in nature may nonetheless be deceptive or may be targeted in a way that discriminates towards particular groups.

6. Support for Technological Advancement

Building projects to understand online public life requires significant domain expertise, technical expertise and start up and operational capital. This is not something that individuals with deep subject matter expertise on topics like information integrity can necessarily do. Our work at Mozilla seeks to reduce barriers to entry in this research space through projects such as Rally.

Growing the overall research ecosystem requires complementary efforts to support and catalyze researchers and domain experts as they invent novel ways of measuring and understanding online life.

This requires far more extensive support from organizations like the NITRD NCO and the NSF. In particular, in addition to the best practices and researcher vetting described above, we encourage you to consider how grantmaking, funding, and programs can help build and maintain measurement systems and platforms capable of supporting the researcher ecosystem. Government engagement on issues like disinformation and threats to information integrity can be fraught, raising challenging issues regarding free expression and speech. Funding support for research platforms however sidesteps these challenges and doesn’t require the government to weigh in on particular disinformation topics. This is a unique, important role that you can play to advance work on this topic.

7. Conclusion

Mozilla is encouraged that the agencies have undertaken the process of reviewing questions pertaining to Federal priorities for research and development efforts to address misinformation and disinformation. It is essential to develop policies and tools to foster transparency and trust in the online ecosystem. The issues addressed in this paper reflect Mozilla’s perspectives and recommendations in key areas. They are not intended to be exhaustive and we would be happy to provide additional detail or further information if helpful.

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

NewsGuard

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.

NewsGuard-IIRD RFI

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

1. *Understanding the information ecosystem:* What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

One of the most significant challenges for researchers investigating the information ecosystem is navigating the opacity of social media platforms' algorithms: specifically how their algorithms curate, downgrade, and deplatform content and users. Until technology platforms provide researchers and industry stakeholders with access to data on the spread of information on their platforms, and the factors that trigger their algorithms to promote or demote content, researchers cannot adequately understand how misinformation spreads online, and thus are hindered in their ability to develop effective strategies to combat the spread.

NewsGuard is an anti-misinformation internet trust service powered by human intelligence, not artificial intelligence, that protects brands, consumers, and democracies from misinformation. Our trained journalists rate the online news and information sources that account for 95% of engagement with the news, assigning each outlet a point score out of 100, a Green or Red rating depending on its score, and a detailed "Nutrition Label" review describing the source and its adherence to standards of credibility and transparency. Together, these nearly 8,000 Nutrition Labels and point scores form our Reliability Ratings dataset of source-level evaluations.

Our second dataset is the Misinformation Fingerprints catalogue — a collection of hundreds of human- and machine-readable myth entries chronicling the most prominent false claims circulating on the internet, spanning topics ranging from the Russia-Ukraine war to COVID-19 misinformation. Each entry contains a description of the myth, a debunk of the myth, citing authoritative sources, variations of the myth, associated keywords and hashtags, and a list of websites we've rated that have perpetuated the myth. This data can be used to seed AI/machine learning tools to track misinformation online at scale, and to aid content moderation efforts on social media platforms, among other use cases.

From our perspective at NewsGuard, having greater access to data from the major platforms would enable us to leverage our expertise to yield insights about how misinformation spreads online and make recommendations about how to mitigate that spread. For example, it would be beneficial to know the proportion of posts shared on platforms like Facebook and Twitter that contain links to news sources that are "Red-rated" and deemed generally untrustworthy by NewsGuard (scoring below 60/100 on our nine apolitical journalistic criteria of transparency and credibility) – or content matching one of our Misinformation Fingerprint entries. Likewise, it would be useful to understand the level of engagement on these posts, the individuals chiefly responsible for perpetuating the misinformation in these posts, and the number of users who were exposed to posts containing such misinformation or content from unreliable sources. It would be of extraordinary value to know what percentage of users of each of the major social media platforms get most of their news and information from unreliable sources—we believe this number is significant and goes to the heart of the information-manipulation crisis affecting many people who are being shown false content and have no tools at their disposal to know which sources are generally reliable and which are not.

Thus far, the major social media platforms have not made these metrics widely available, thus hindering important research in the field of misinformation. Without this level of transparency, research practitioners are unable to adequately assess the full extent of the problem.

A related challenge facing researchers in this field is determining how to define the scope of "news" or "information" online. Defining the online information ecosystem has become even more complex as a result of the fact that platforms host user-generated content while also retaining algorithmic and curatorial authority over what users ultimately see on their platforms. The industry needs a baseline definition of what constitutes a news or information source to not only make disparate research efforts in the field more comparable, but also to hold platforms to account by requiring them to clearly delineate the parts of their business that publish news and authoritative information from those that publish user-generated content. Further, industry stakeholders need a common understanding of what distinguishes a reliable news and information source from an unreliable one.

Conceptualizing such a definition of news and information sources is the collective responsibility of regulators, platforms, and industry, and it is a decision that NewsGuard is well-positioned to help inform. For example, NewsGuard has already become the industry standard for source reliability classifications among academic researchers studying misinformation, used by top researchers at Stanford, Dartmouth, New York University, Northeastern University, the German Marshall Fund, and other institutions. Our work is used by entities from the World Health Organization to the Pentagon's Cyber Command and the State Department's Global Engagement Center.

In sum, researchers, publications, news aggregators, platforms, and content moderators would benefit from both (1) a baseline definition of what constitutes a news or information source that is transparently communicated to all stakeholders and (2) an impartial, objective third-party source reliability assessment to use as a common denominator when promoting and curating information for the benefit of consumers and quality newsrooms alike.

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. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?

When it comes to the investigation and identification of information manipulation, much attention has been paid to developing AI-based solutions that can track the provenance and spread of dubious content at scale. One example is the Content Authenticity Initiative (CAI) announced by Adobe in 2019 in partnership with Twitter and the New York Times.[1]

But such technology-based solutions are incomplete if they are not coupled with the expertise of human experts capable of parsing the subtleties of misinformation. More focus must be placed on developing human-intelligence solutions that can be paired with AI technologies in the process of information investigation. Early research conducted by NewsGuard with AI technology company Blackbird AI shows that NewsGuard's source Reliability Ratings and Misinformation Fingerprints library of false content can augment the efficacy of AI-based misinformation tracing efforts.[2] More research should be done to extend these early attempts to combine human and artificial intelligence in pursuit of holistic and effective misinformation detection.

In terms of mitigation efforts, strategies for protecting information integrity are generally most sustainable when they elicit lasting behavioral change among news consumers. Large platforms and companies with large audiences should have a responsibility to provide user empowerment tools that equip consumers with greater context and authoritative information in order to make their own decisions, centering the user

[1] https://contentauthenticity.org

[2] https://www.newsguardtech.com/wp-content/uploads/2020/10/NewsGuard-x-Blackbird.AI-Joint-Report-1.pdf

experience instead of censoring information or blocking content. One example of such a user empowerment intervention is NewsGuard’s browser extension, which displays Red and Green shields next to links to sources, indicating the general reliability of a source at the moment a user first encounters a piece of content. From there, users can read the full “Nutrition Label” review of the source, empowering them to gain richer context behind a source without blocking them from accessing any content. Researchers and other industry stakeholders should promote research and development of tools like NewsGuard, which is recommended by the European Commission to the digital platforms to help them comply with the user-empowerment requirements of the European Commission Code of Practice on Disinformation and by the UK government as it has identified online safety tools the platforms can provide their users.

3. *Information awareness and education:* A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

Forthcoming research should focus on evaluating the efficacy of educational interventions on users’ ability to discern information manipulation, such as NewsGuard’s browser extension, which provides seamless source reliability ratings alongside social media posts and search results. Research that demonstrates the effectiveness of equipping online users with contextual information and source credibility indicators would help regulators and educators implement tools for students and young people to become more discerning consumers of online information.

For example, researchers and practitioners might look to NewsGuard’s partnership with the American Federation of Teachers, one of the largest teachers’ unions in the U.S. Through the partnership, the AFT’s 1.7 million educator members receive free access to the NewsGuard browser extension for their 20 million students and their families, providing a unique media literacy intervention. Studies that test the efficacy of NewsGuard and similar tools in classrooms would accelerate the adoption of these research and contextual information tools.

Similarly, NewsGuard integrates with new media companies that are reimagining existing platforms and offering user-centric benefits like privacy, ads-free interfaces, and source reliability context. Bright, a new social media app that emphasizes user privacy and experience, displays NewsGuard’s ratings within the app where news links are posted to give users and moderators alike greater context for the content being shared.[3] Neeva, an ads-free subscription search engine founded by former Google and YouTube executives, similarly displays NewsGuard data in its search results.[4] Further research measuring the effectiveness of NewsGuard integrations in social and search platforms would help make the case for other technology companies to incorporate NewsGuard's user empowerment tools to make the online media environment a safer place.

4. *Barriers for research:* Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

One of the central barriers to conducting information integrity R&D is insufficient baseline context for the credibility of sources and individual pieces of content in the sprawling online information ecosystem. Presented with a Twitter dataset of billions of tweets, for example, a researcher is tasked with the challenge of determining which pieces of content may contain misinformation.

NewsGuard offers one solution to this quandary by providing human-curated, independent assessments of the reliability of news sources, providing researchers with "ground truth" for quickly determining which pieces of content contain information from dubious sources as opposed to reliable sources. For example, researchers at the University of Michigan use NewsGuard ratings as sources for their regular "Iffy Index," which measures the quantity of misinformation on the various digital platforms over time. Moreover, NewsGuard's "Misinformation Fingerprints" dataset, which offers a constantly-updated catalog of the top myths spreading online, can be used to match individual pieces of content to specific known false claims. By combining both datasets together, researchers can pinpoint content containing falsehoods with even greater accuracy.

Having licensed our data to dozens of academic and government researchers studying misinformation, NewsGuard would welcome the opportunity to provide its data to a greater number of independent researchers under a wider, government-funded license.

5. *Transition to practice:* How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

[3] https://www.newsguardtech.com/press/newsguard-partners-with-bright-the-new-ethical-social-media-app-to-keep-misinformation-off-the-platform/

[4] https://neeva.com/press/neeva-and-newsguard-team-up

Federal government support for turnkey solutions like NewsGuard's could equip tens of thousands of users with tools to evaluate the reliability and credibility of news websites while browsing the search engine results, social media feeds, or the open internet. NewsGuard is a ready-to-deploy solution with a proven record in schools, universities, public libraries, search engines, and social platforms.

NewsGuard's Red and Green source ratings, trust scores out of 100, and detailed Nutrition Label reviews all serve to inform users as they navigate the internet. As a browser extension, NewsGuard does not interfere with user behavior or block content: instead, it seamlessly provides source level reliability ratings alongside the browsing experience to help users decide whether they want to trust a certain news source.

A wealth of research already supports the effectiveness of NewsGuard's intervention:

- After launching the browser extension, NewsGuard and the Knight Foundation commissioned a Gallup survey to assess how the tool worked when installed on personal computers. The study, conducted in November 2018, found that 91% of respondents found the NewsGuard Nutrition Labels helpful, and 90% generally agree with the ratings and respondents trusted the ratings more because NewsGuard ratings are done by "trained journalists with varied backgrounds." The Gallup researchers concluded: "The positive results among people who accepted Gallup's invitation to download the NewsGuard browser extension suggest a desire for more information about the sources of news people see online, such as in their social media newsfeeds and in their search results. The news source rating tool offers a scalable solution to identify which news sources adhere to the basic journalistic standards of accuracy and accountability citizens expect and deserve."

- In 2017, Indiana University researchers Alan Dennis and Antino Kim ran an investigation comparing different reputation rating formats to assess their ability to influence users' belief of news articles.[5] They found that presenting source reputation ratings directly influences the extent to which users believe articles on social media. In a 2019 article in The Conversation,[6] Dennis and Kim said: "What we learned indicates that expert ratings provided by companies like NewsGuard are likely more effective at reducing the spread of propaganda and disinformation than having users rate the reliability and accuracy of news sources themselves."

[5] https://misq.org/says-who-the-effects-of-presentation-format-and-source-rating-on-fake-news-in-social-media.html

[6] https://theconversation.com/rating-news-sources-can-help-limit-the-spread-of-misinformation-126083

- In their article “Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence,” published in May 2017 in PLOS One, Stephan Lewandowsky, a psychologist at the University of Bristol, John Cook, a researcher at the Center for Climate Change Communication at George Mason University, and Ullrich Ecker, a cognitive psychologist at the University of Western Australia, explain how, because pre-existing beliefs impact how people respond to novel information, warnings about misinformation are more effective when they are administered before misinformation is encountered rather than after.[7] NewsGuard’s intervention of providing source-level evaluations when someone first encounters content can be described as “pre-bunking,” supporting the findings of Lewandowsky et. al’s study.

6. *Relevant activities:* What other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

Social media and technology platforms with significant consumer bases and audiences have a duty of care to protect their users from the harms of mis- and disinformation, according to the UK Online Safety Bill.[8] The Online Safety Bill aims to strengthen the UK’s stance on internet regulation by holding large tech companies to account for protecting users from online harms and keeping harmful information off their platforms, with major fines and regulatory consequences for firms that fail to comply. An exemplar of a private company that has proactively addressed such harms across its product areas is Microsoft, which has forged a global, company-wide partnership with NewsGuard to safeguard employees, clients, and users from misinformation. Teams across Bing, MSN, Teams, Edge, and the Democracy Forward Initiative have relied on NewsGuard’s trustworthiness indicators for news and information websites, setting a precedent for other technology platforms to forge proactive partnerships that safeguard information integrity for users. The U.S. Federal Government is well positioned to capitalize on NewsGuard’s solution for both internal research purposes and broader licensing to agencies and associated companies, following the success of NewsGuard’s work with private sector stakeholders like Microsoft, as well as public sector researchers and analysts at groups like the Pentagon.[9]

[7] https://pubmed.ncbi.nlm.nih.gov/28475576/

[8] https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament

[9] https://www.newsguardtech.com/press/newsguard-wins-pentagon-state-department-contest-for-detecting-covid-19-misinformation-and-disinformation/

In Europe, the European Commission is currently convening dozens of signatories for the revised Code of Practice on Disinformation, a self-regulatory instrument that enables technology platforms to collectively set common goals for addressing misinformation and associated KPIs for tracking progress, with self-reporting of progress. A section on “empowering users” outlines how the platforms should provide tools to their users from independent entities using apolitical, transparent criteria to alert them to news sources that publish misinformation. It is the first time that global industry stakeholders have agreed, on a voluntary basis, to self-regulatory standards to fight disinformation by setting a wide range of commitments, from transparency in political advertising to the closure of fake accounts and demonetization of purveyors of disinformation.[10] As the U.S. moves closer to formalizing legislation that holds technology platforms accountable and protects internet users from online harms, the EU’s self-governing frameworks are a useful model to consider.

There also exists the opportunity to consider legislative mechanisms that promote media literacy education. Numerous U.S. states have recently considered or enacted different laws requiring media literacy education in schools, with Illinois being the first state to require a unit of media literacy education as a prerequisite for graduating high school.[11] Fourteen other states currently have some standards for media literacy, but none require a unit of instruction. Elevating and enforcing standards for media literacy education across the country — and providing resources like NewsGuard to do so — would yield more informed and discerning online citizens for generations to come.

7. *Support for technological advancement:* How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission:

- **Ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security; and**
- **maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.**

The European Union through its Code of Practice on Disinformation and the UK through its Online Safety Bill have taken the lead on addressing misinformation by encouraging digital platforms to provide news consumers with news-literacy tools. Many online safety tools have been developed in the U.S., and the White House Office of Science and Technology can add its support to this growing industry, including by building

[10] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

[11] https://www.chicagotribune.com/opinion/commentary/ct-opinion-illinois-bill-media-literacy-fake-news-schools-20220228-p5q2xcuj3zautaedsi4hqfadga-story.html

awareness among the platforms of the benefits to them and their users from making third-party “middleware” tools available to consumers.

Ratings of news sources at the domain level has been proven an effective tool for news consumers, if these ratings are done in an open, disclosed and transparent manner, using criteria that are objective and apolitical. Restoring trust to generally trustworthy sources of news and information supports democratic values and democratic systems. A free and trustworthy press has been at the center of the American experiment since its beginning and must be restored if our democratic system is to be as strong and resilient as the people expect.

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

Partnership for Countering Influence Operations, Carnegie Endowment for International Peace (PCIO-CEIP)

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.

Subject: RFI Response: Information Integrity R&D

Date: 2 May 2022

To: NCO

From: Researcher from The Partnership for Countering Influence Operations, Carnegie Endowment for International Peace

Protecting information integrity in any substantial way first requires a comprehensive understanding of the information environment. The information environment is the space where humans and machines make sense of the world using information, which can be anything from ideas to words to videos. This information moves through a complex and dynamic system via channels like television and digital media, but also through people face-to-face. Disparate fields study aspects of the information environment but lack a systematization of any resulting knowledge. As this system increases in complexity, so too do challenges for understanding it, as the demands for engineering resources become more costly and thereby scarce for academic and policy research. Significant resources are urgently required to advance understanding of the information environment and threats within it, specifically through investments for shared, multistakeholder scientific infrastructure to power policy-relevant research.

WHAT ARE THE GAPS?

Little is known about how the information environment works as a system. Most research on threats within the information environment, such as disinformation, focuses on specific examples of activity often in isolation from the audience it targets or the ecosystem in which it spreads. Put another way, researchers know little about the flow of information between different types of online platforms, as well as through influencers, and into other types of media. Furthermore, research tends to focus on single platforms, most often Twitter, whose monthly active users are fewer than platforms such as Snapchat.[1]

Knowledge about the effects of influence operations also remains limited. Most research on the effects of influence operations focuses on traditional mass media.[2] Where social media is the focus of measuring the effects of disinformation, most work explores short-

[1] Statista Research Department, "Global social networks ranked by number of users 2022", *Statista*, (March 8, 2022). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2] Courchesne, Laura, Jacob N. Shapiro, and Isra M. Thange. “Review of Social Science Research on the Effects of Influence Operations,” *Carnegie Endowment for International Peace*, (2021). https://carnegieendowment.org/2021/06/28/measuring-effects-of-influence-operations-key-findings-and-gaps-from-empirical-research-pub-84824

term effects only. Little is known about the longer-term effects of disinformation conducted over social media although some of the top platforms today launched almost 20 years ago. How does engagement with digital media change user beliefs or views over time? Does repeated exposure to anti-vaccination disinformation make people more or less likely to get inoculated, and how long do those views last? Are there other impacts from such exposure, like a reduced trust in media or government, and what do those potential shifts mean for civic engagement, such as voting or compromising with others who hold different views?

Even less is known about the efficacy of common countermeasures.[3] Most research assessing the impact of interventions focuses on disclosures by social media companies to users potentially exposed to influence operations, fact-checking, and content labeling. Results are promising for fact-checking as an immediate countermeasure, reducing the impact of false information on beliefs as well as on subjects' tendency to share disinformation with others. However, very few researchers have studied interventions by social media companies, such as deplatforming, algorithmic downranking, or content moderation. While social media platforms have introduced new policies and experimented with interventions to counter threats like disinformation, there is little publicly available information on whether those measures work—in fact, one count from early 2021 found that platforms provided efficacy measurements for interventions in only 8 percent of cases.[4]

Part of the problem is a lack of access for researchers to data held by private companies. Yet, even if data-sharing challenges could be solved, the fact that academic research is seldom designed with policy development in mind means that critical learnings are not being translated and used systematically to support evidence-based policy development. Moreover, conducting measurements related to disinformation, especially over time, requires the cooperation of the platform studied, if for no other reason than to keep the research team informed of changes made to it that could affect the research. This suggests that data access alone will not solve this problem, but also a mechanism to facilitate independent research with access to researchers working inside the platforms is necessary too.

The research community's approach to producing work is also wildly inefficient. Each project runs its own small data processing pipeline, not designed with an eye to re-useability across multiple studies. Often these studies are conducted by graduate

[3] Courchesne, Laura, Julia Ilhardt, and Jacob N. Shapiro. "Review of social science research on the impact of countermeasures against influence operations". *Harvard Kennedy School (HKS) Misinformation Review* (13 September 2021). https://doi.org/10.37016/mr-2020-79

[4] Bateman, Jon, Natalie Thompson and Victoria Smith. "How Social Media Platforms' Community Standards Address Influence Operations," *Carnegie Endowment for International Peace*, (01 April 2021) https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201 and Yadav, Kamya. "Platform Interventions: How Social Media Counters Influence Operations," *Carnegie Endowment for International Peace*, (25 January 2021) https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698

students or post-docs with little training or experience. In effect, the status quo entails future social scientists building bespoke data science pipelines, instead of doing social science research. This lack of proven research designs puts the community perpetually behind the curve in understanding things such as the impact of interventions to disinformation.

WHAT NEEDS TO BE DONE?

REVISIT FUNDING MODELS AND COLLABORATION

Stable funding and more coordination between stakeholders are required to counter threats within the information environment. The field is fragile and fragmented. Funding tends to be project-based, meaning most initiatives lack the resources to fund ongoing operations.[5] Likewise, the need to show results and success to donors encourages initiatives to overinflate their effectiveness as they seek continued funding. While the disparate community of researchers increasingly connects through professional convenings, collaboration between experts is still ad hoc. For example, experts don't tend to build on each other's work, as demonstrated by a lack of citations between them in drafting similar policy recommendations. Between sectors there are significant disconnects, with academics, civil society, industry, and government often operating within their own communities, diminishing their ability to understand each other and their respective roles.[6]

A CERN FOR THE INFORMATION ENVIRONMENT

Longer-term research collaboration must be facilitated. A permanent mechanism is needed for sharing data and managing collaborations that ensures the independence and credibility of researchers. Given the costs associated with and expertise required to conduct measurements research, an international model whereby several governments, philanthropists, and companies cooperatively support an independently governed research enterprise is needed.

One model is a multi-stakeholder research development center, which would enable cross-sector collaboration with an emphasis on facilitating information-sharing to test hypotheses and design countermeasures and intervention strategies.[7] Another follows

[5] Smith, Victoria and Natalie Thompson. "Survey on Countering Influence Operations Highlights Steep Challenges, Great Opportunities," *Carnegie Endowment for International Peace*, (7 December 2020) https://carnegieendowment.org/2020/12/07/survey-on-countering-influence-operations-highlights-steep-challenges-great-opportunities-pub-83370 and Yadav, Kamya. "Countering Influence Operations: A Review of Policy Proposals Since 2016," *Carnegie Endowment for International Peace*, (30 November 2020) https://carnegieendowment.org/2020/11/30/countering-influence-operations-review-of-policy-proposals-since-2016-pub-83333

[6] Yadav, Kamya. "Countering Influence Operations: A Review of Policy Proposals Since 2016," *Carnegie Endowment for International Peace*, (30 November 2020) https://carnegieendowment.org/2020/11/30/countering-influence-operations-review-of-policy-proposals-since-2016-pub-83333

[7] Shapiro, Jacob N., Natalie Thompson, and Alicia Wanless. "Research Collaboration on Influence Operations Between Industry and Academia: A Way Forward," *Carnegie Endowment for International Peace* (3 December 2020)

the European Organization for Nuclear Research, or CERN, model, supported by several countries working with multiple research organizations to create an international network, thus fostering a wider field.[8] This could include a fellowship model where multiple stakeholders including members from platforms, government, academics, and civil society organizations, come together to propose a specific policy-related solution. An international approach would help address the imbalance in available research that skews toward the Global North, thus helping those in countries where malign influence operations are often a matter of life and death. Another benefit of an international model is that it bakes in a fail-safe should any single member country slide into autocracy, ensuring that such a center cannot be abused.

Such an institute could support the research community studying how the modern information environment is impacting society, by building shared scientific infrastructure. An institute offering shared engineering infrastructure, could fill gaps such as collecting representative samples across multiple countries, so that scholars can quickly test their ideas in many settings, without having to learn each anew. Or it could create standardized datasets of labeled posts, which would speed development of tools to make sense of the conversation.

Finally, in systematizing research, such an institute could lead on the development of theory that helps connect myriad fields applying their respective methods to analyzing aspects of the information environment, establishing an information ecology. Such an institute could act as the center of gravity to articulate definitions and frameworks for understanding the information environment, providing much needed foundational knowledge and consilience, and overcoming yet another persistent challenge in protecting information integrity – speaking a common language about the problem.

https://carnegieendowment.org/2020/12/03/research-collaboration-on-influence-operations-between-industry-and-academia-way-forward-pub-83332

[8] Lewandowski, Stephan, Laura Smillie, David Garcia, Ralph Hertwig, Jim Weatherall, Stefanie Egidy, Ronald E. Robertson, Cailin O'connor, Anastasia Kozyreva, Philipp Lorenz-Spreen, Yannic Blaschke, and Mark Leiser. "Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making." *JRC Science for Policy Report*, European Commission, (2020) https://publications.jrc.ec.europa.eu/repository/handle/JRC122023 and * Wanless, Alicia. "What's Working and What Isn't in Researching Influence Operations?, *Lawfare*, (22 September 2021). https://www.lawfareblog.com/whats-working-and-what-isnt-researching-influence-operations

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

Patrick W.

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.

From: Patrick W.

Subject: RFI Response: Information Integrity R&D

Date:

Hi. Thanks for soliciting comments.

To me there are 2 issues, and clear solutions to both. The problems:

1. Misinformation / disinformation / propaganda on public media (internet, TV, radio, public journalism, newspapers, etc.)
2. Sensationalization of news (as opposed to educating the public and serving the public good)

The solutions:

1. A constitutional amendment specifying that all information provided on publicly regulated channels (all RF, cable and internet) and news services must be true, verifiable and not presented in a misleading way. This must be supported by federal law specifying criminal and civil penalties for violations that would always exceed any income gained through such violations. (Wording could be similar to the laws requiring advertisers to be honest.)
2. Federal law requiring all public channels which either purport or appear to be news sources (e.g., any channel, network or URL with the term *news* in its name) to give priority to news items that are in the public interest and to present both sides of controversial issues. The wording here should say something about the necessity for an informed electorate and the responsibility of public channels to help the public understand issues in a truly holistic way. Opinion pieces must be clearly labeled as such, and presented after the more objective and factual information.

Democracy and individual liberty are hard work. Propaganda and echo chambers make it practically impossible. We must do everything we can to give everyone access to real facts and honest analysis and protect the public from disinformation and propaganda. The unity and survival of our democratic republic depends on it.

Thank you for your hard work and integrity.

Patrick W.

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

Ripeta

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.

Ripeta RFI Response

Prepared for: **NITRD NCO and NSF**

Prepared by: **Researchers and the Ripeta Team**

Ripeta, part of Digital Science
www.ripeta.com

About Ripeta:

Ripeta believes preserving the integrity of information is of the highest importance to ensure our society is protected against information manipulation. We are excited to respond to this RFI in support of the Information Integrity Research and Development Interagency Working Group (IIRD IWG) as a core part of our mission is to develop appropriate tools in accord with efforts to improve integrity and quality of research and development.

Ripeta services have grown organically from the research ecosystem, in an environment where government agencies, researchers, publishers and funders are producing research at an increasingly rapid pace. Ripeta services can and will specifically empower governmental staff to increase their efficiency, while ensuring high levels of accuracy in monitoring, analyzing and reporting on author integrity and other critical aspects of quality research publication.

Ripeta was founded in 2017 by Dr. Leslie D. McIntosh and Cynthia Hudson Vitale, and has the primary mission to assess, design, and disseminate practices and measures to improve the reproducibility of science with minimal burden on scientists and the agencies and organizations in which they work.

Ripeta focuses on assessing the quality of reporting and the robustness of the authors and scientific method by offering quality checks around the following quality indicators:

- Author Identification
- Author Contribution Statement
- Count of Authors (single)
- Competing Interests
- Ethical Approval Statement
- Ethical Approval Organisation
- Funding Source
- Funding Grant ID
- Data Availability Statement (DAS)
- Data Locations
- Code Availability Statement
- Analysis Software
- Others available upon request

By improving the reproducibility of science through automation, Ripeta aims to make better science by identifying and highlighting the important parts of research that should be transparently presented in a manuscript and other materials. Our automated reproducibility assessment tool detects and predicts the reproducibility of scientific research, improving evidence-based science and fiscal efficiency of research investments.

We understand that protecting the integrity of the information ecosystem requires an understanding of actors and consumers of information and our technology and services were designed and developed for this exact purpose.

Our technology can detect and mitigate information manipulation across a wide range of information media and forms, and in combination with Altmetric solutions, we can monitor and report activity levels across various communication modalities.

Information Requested:

Digital Science and Ripeta can assist the IIRD IWG on its priorities for information integrity research and development (R&D). Below are our responses to the seven (7) areas of IIRD IWG focus.

1. Understanding the information ecosystem: There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem.

Q: What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?

A: Key research challenges in providing a common foundation for understanding information manipulations include:

- **Lack of diversity.** Funding a diverse group of people must occur for robust understandings of and solutions for countering and preventing information manipulation to be successful.
- **Insufficient collaboration.** Funding solutions from multi-sectors (e.g. academia, small-business) will foster greater options to counter the problems with information manipulation.
- **Limited views of global research collaborative networks.** Understanding and mapping the ways in which research collaborations develop and evolve over time can help identify publishing patterns in various research domains, and perhaps lead to earlier identification of targeted information manipulation.

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.

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

A: Prioritizing Technological Solutions. The key to success for any organization in reducing fraud and data manipulation is to deploy data analysis and monitoring technology across a variety of reporting system touchpoints. These technologies, when deployed and adjusted to the proper sensitivities, will automatically trigger alerts when data is in contrast to expected results and parameters. A comprehensive gap analysis is required to determine how best to deploy technologies.

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

A: Knowledge of impact on decision-making as it pertains to social, economic, environmental, and health outcomes. For example, when Indigenous groups have not been included in research, outcomes, and knowledge transfer with and from the perspective of Indigenous individuals and groups, it is difficult to understand and mitigate how information has been manipulated with the potential to harm community outcomes.

3. Information awareness and education: A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information.

Q: What issues should research focus on to understand the barriers to greater public awareness of information manipulation?

A: From the perspective of a technology company, two of factors that we believe researchers should focus on to understand these barriers are:

1. Wise and consistent implementation of FAIR data sharing policies.
2. Use of technology to quickly identify trends and create transparency around information shared, particularly trending information. An area of particular concern is to understand how much information and awareness is enough and how much is too much. A tactic of information manipulation is to overwhelm with information, causing decision fatigue.

Q: What challenges should research focus on to support the development of effective educational pathways?

A: Enabling the public to easily gain proper context around shared information and facilitating simple methods of filtering out known, demonstrable falsehoods. Educational pathways are most effective when based on Trust and Integrity. This is the core of Ripeta's mission.

4. Barriers for research: Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise,

data, and practices across the full spectrum of stakeholders, both domestically and internationally.

Q: What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

A: Please see our responses above.

5. Transition to practice:

Q: How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

A: By utilizing the most advanced research integrity monitoring, supporting both educational and solution driven approaches, and supporting the maintenance of such solutions.

6. Relevant activities:

Q: What other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

A: Digital Science invests in collaborative partnering with the G20 and other leading government agencies across the world to promote and facilitate greater rigor throughout the global research ecosystem. As a governmental partner and with a focus on FAIR principles, we are eager to share the global trends our extensive data uncovers. In this regard, we have experts in many key areas of information integrity that will inform and feed into the US Federal information integrity R&D strategic planning. As a technology company, Ripeta and Digital Science are best able to shed light on technology solutions that should be considered. Some of our leading strategic activities include: author identification and verification, the use of artificial intelligence to detect nefarious activities of individuals and groups, and tracking down and reporting global paper mill activities.

7. Support for technological advancement:

Q: How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission:

- **Ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security; and**
- **Maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.**

A: Ripeta's mission and core-values along with the mission of Digital Science and each of its portfolio companies are very closely aligned to the mission of the White House Office of Science and Technology. Please see the information in these links for further detail:

- https://www.digital-science.com/challenge/knowledge-creation/
- https://www.digital-science.com/challenge/insights-that-support-decisions/

The U.S. Federal governmental agencies should aim to:

- ***Save time*** by automatically checking research for quality indicators up front
- ***Produce higher quality research papers*** by filtering out those papers that don't have the components required for reproducibility
- ***Underpin reputation management*** while increasing public trust in research by avoiding retractions due to authorship issues
- ***Compare agencies across time*** to themselves and other agencies in order to better understand how indicators of research quality compare

Ripeta is part of Digital Science & Research Solutions Ltd. (Digital Science)

We are grateful for the opportunity to share insights in response to this RFI.

Sincerely,

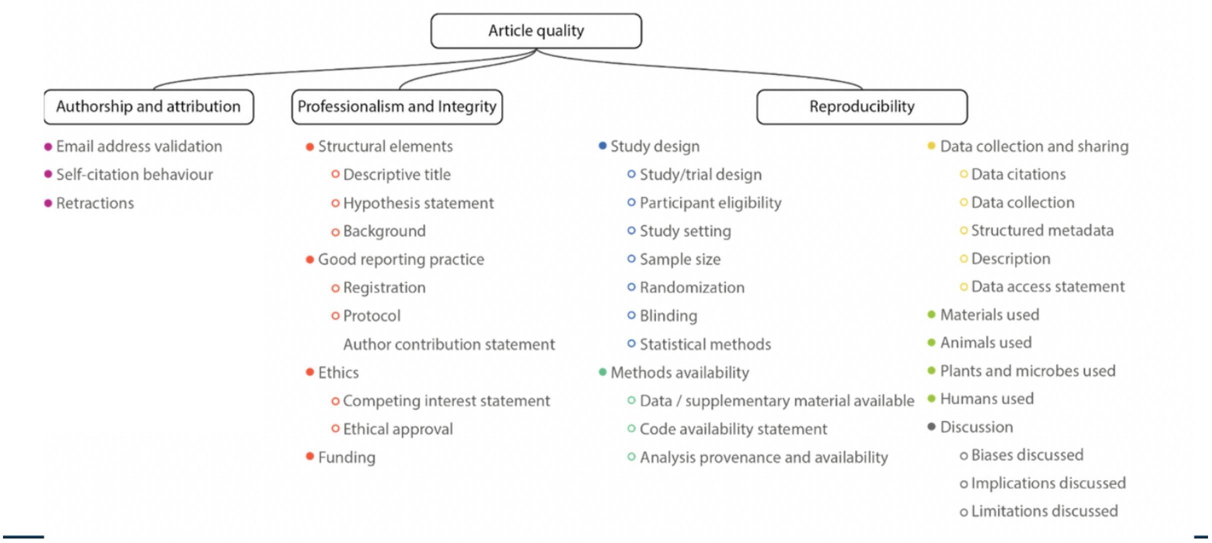
Ripeta: Business Development

SUPPLEMENTAL Information:

Accuracy of Measurements

The Ripeta Data Science team develops data models using artificial intelligence and machine learning to examine published literature and give scores to reflect aspects of research integrity. Examples of the factors considered are shown below..

Taxonomy to Trust



Article quality

Authorship and attribution

- Email address validation
- Self-citation behaviour
- Retractions

Professionalism and Integrity

- Structural elements
  - Descriptive title
  - Hypothesis statement
  - Background
- Good reporting practice
  - Registration
  - Protocol

  Author contribution statement
- Ethics
  - Competing interest statement
  - Ethical approval
- Funding

Reproducibility

- Study design
  - Study/trial design
  - Participant eligibility
  - Study setting
  - Sample size
  - Randomization
  - Blinding
  - Statistical methods
- Methods availability
  - Data / supplementary material available
  - Code availability statement
  - Analysis provenance and availability
- Data collection and sharing
  - Data citations
  - Data collection
  - Structured metadata
  - Description
  - Data access statement
- Materials used
- Animals used
- Plants and microbes used
- Humans used
- Discussion
  - Biases discussed
  - Implications discussed
  - Limitations discussed

Related Reading: Blogs, Reports, Surveys, and Use Cases

Research Square Launches Beta Testing of Ripeta's Open Science Assessment Tool:
https://ripeta.com/research-square-launches-beta-testing-of-ripetas-open-science-assessment-tool/

Imposters and Impersonators in Preprints: How do we trust authors in Open Science?:
https://scholarlykitchen.sspnet.org/2021/03/17/imposters-and-impersonators-in-preprints-how-do-we-trust-authors-in-open-science/

Wellcome and Ripeta partner to assess dataset availability in funded research:
https://ripeta.com/wellcome-and-ripeta-partner-to-assess-dataset-availability-in-funded-research/

Trusting Science in the Time of Coronavirus:
https://ripeta.com/trusting-science-in-the-time-of-coronavirus/

The Anatomy of a Data Availability Statement (DAS):
https://ripeta.com/the-anatomy-of-a-data-availability-statement-das/

Ripeta Background:

Founded: In 2017 by Dr. Leslie D. McIntosh, PhD (CEO) and Cynthia Hudson-Vitale, MA, after working together at Washington University Medical Center and the University Libraries (St. Louis, Missouri U.S.A.), founded Ripeta.

Mission: Improve professionalism, reproducibility, and integrity of scientific research

Delivery: Utilize technology and automation to improve accuracy, save time, efforts and costs, while minimizing burdens on publishers, researchers and funders.

Ripeta's mission is to assess, design, and disseminate practices and measures to improve the reproducibility of science with minimal burden on scientists. We focus on assessing the quality of reporting and the robustness of the authors and scientific method. At Ripeta we believe that the credibility of a paper can be boiled down to three indicators of trust: research, professionalism, and reproducibility. To perform well in all three of these areas, the author must include specific elements which bolster the integrity of their manuscript.

Trust in the Research intends to determine whether papers reviewed deserve the general specification of a 'research' article.

Trust in Professionalism is based around ascertaining the legitimacy of the authors and whether their reporting adheres to established standards of professionalism.

Trust in Reproducibility is centered around the elements of a paper which may facilitate a future researcher's ability to achieve the same results when replicating the original study.

Ripeta offers many solutions and services to increase the integrity of research.

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

Sandia National Laboratories

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.

SANDIA REPORT
SAND2022-6477 O

Sandia National Laboratories

Response to RFI on Federal Priorities for Information Integrity Research and Development

Prepared by
Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information

U.S. Department of Commerce
National Technical Information Service

1. INTRODUCTION

This document serves as Sandia National Laboratories' (i.e., Sandia's) response to the Request for Information (RFI) on Federal Priorities for Information Integrity Research and Development. The RFI defines information integrity preservation as protecting society against information manipulation and defines information manipulation as, "activities that aim to influence specific or multiple audiences through disinformation, misinformation, malinformation, propaganda, manipulated media, and other tactics and techniques that intentionally create or disseminate inaccurate, misleading, or unreliable information." [19]

Successfully preserving information integrity requires integrating a broad range of perspectives, including political, policy, social, and scientific. This document's contribution is to address the problem from the research and development (R&D) perspective of a national laboratory. Our perspective involves strategically planning a broad R&D program to support national security. This requires R&D along the entire information ecosystem using a scientifically grounded approach. This approach treats information manipulation as an object of scientific study, develops an understanding of that object, and then creates new technology to help solve the relevant national security problems.

We address two questions from the RFI. A scientifically grounded approach must define the phenomena being studied, so we address Question 1: "Understanding the information ecosystem." We describe the information ecosystem by defining a model the federal government can use to help organize research and development activities. Much of our national security focused R&D is aimed at preserving the integrity of the systems we support, so we also address Question 2: "Preserving information integrity and mitigating the effects of information manipulation." We then describe the points in this model that are vulnerable to information manipulation, which we refer to as the attack surface. We include a sample of relevant research, including research at Sandia. While not comprehensive, this sample illustrates the kinds of research that can help address this critical issue.

2. INFORMATION ECOSYSTEM

Addresses Question 1: Understanding the information ecosystem

Models of information integrity have been developed and discussed since at least the early 1970s [16]. The information ecosystem has grown exponentially in its complexity with the widespread availability of cheap computation and communication. The discussion of information integrity continues today (e.g. [13]).

Claude Shannon, the founder of Information Theory, famously wrote, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected

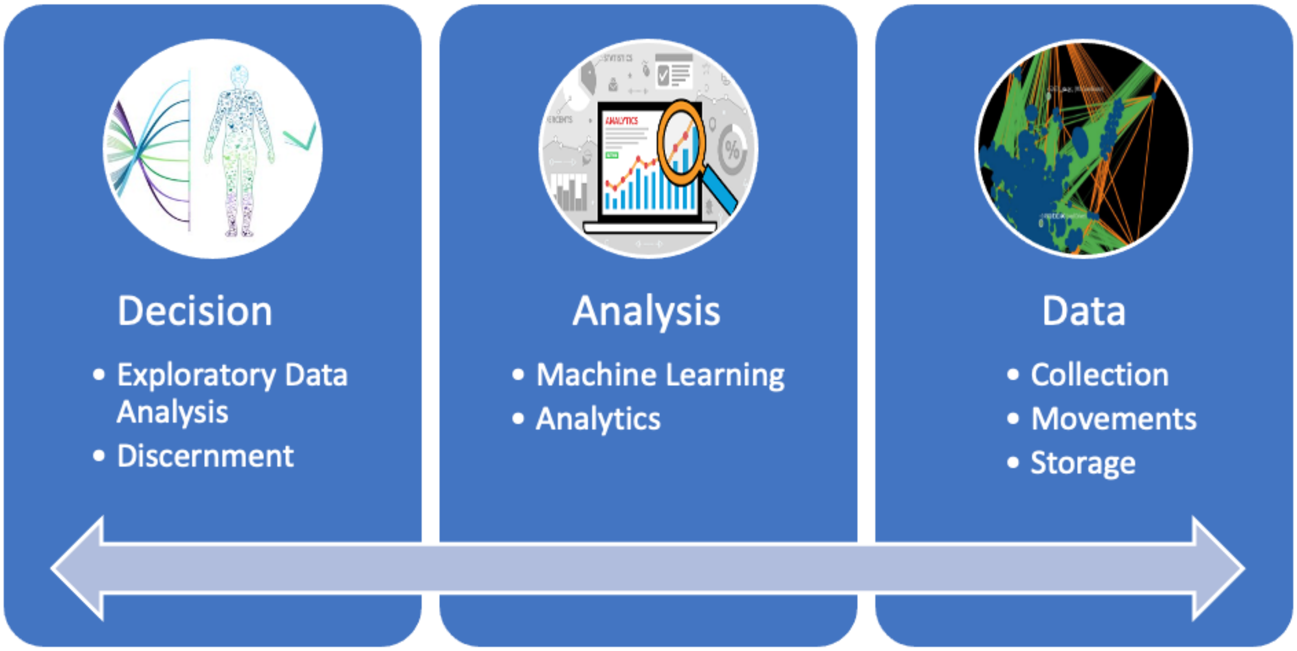


Figure 2-1. Model of the information ecosystem.

at another point," and that the "semantic aspects of communication are irrelevant to the engineering problem" [20]. This description presents two aspects of any information ecosystem: the "engineering problem" and the "semantic" problem. The "engineering problem" is simply about storage and communication of bits independently of their meaning. The "semantic aspects" that Shannon saw as irrelevant to the engineering problem are critical to the issue of information integrity. We believe that addressing both the semantics and the engineering problem is critical to the protection of information ecosystems. Because so much of today's technology is rooted in Shannon's theory, we also believe that this also presents promising avenues for R&D investment.

Figure 2-1 shows the model we use for understanding the information ecosystem. This model accounts for the full spectrum of the information ecosystem, from bits to semantics. It has three categories, each dealing different aspects of information and raising different issues related to information integrity. These categories will be described in detail in the next section. We believe this model is well suited for defining information integrity risks and planning R&D activities in support of national security.

2.1. Three Categories of the Information Ecosystem

The first category of the information ecosystem is Decision. This category is especially focused on the semantic aspects of information, including its context and its intended use. One key risk in this area is losing the context and assumption(s) made when collecting data. Loss of context and assumptions opens the information ecosystem to compromise. Another key risk is disinformation. Disinformation, defined as intentionally leading people to false beliefs or conclusions [28], is a threat to information integrity. In addition to detecting disinformation, research opportunities in this category include issues such as culture and the conditions under which people would find it acceptable to lie [27]. R&D in this category is especially focused on the people creating and consuming the data. Because of this, successful R&D in this category requires the incorporation of psychological research about how people respond to the engineered systems with which they interact.

The second category of the information ecosystem is Analysis. We define this category as the augmentation of information through algorithms. Whereas the Decision category is about the meaning of the information, the goal of the Analysis category operates consistently with semantics, but the algorithms have no inherent understanding of the underlying meaning because they generally accomplish complex pattern matching without broader context. For example, an image classifier that can find images containing tanks does not encode what a tank is, what it is used for, or why a decision-maker would need to find one in an image. This creates risk to the information ecosystem.

The third category is Data. Shannon's "engineering problem" is the chief object of the Data category. The key concern is ensuring that the bits themselves are accurately stored, transmitted, and reproduced, independent of their semantics. This category includes cybersecurity, encryption, authentication, authorization, and robustness of data, along with the engineering around how that data is visualized and presented to users. The risk to the information ecosystem here is broad and well understood, but there is still room for R&D. The internet and related systems were originally developed for speed and robustness at the expense of security. This remains an unsolved problem, creating a variety of R&D needs.

These three categories are highly interrelated and thus must be addressed by interdisciplinary teams. The Analysis category presupposes the existence of data, and the Decision category assumes the existence of data augmented by analysis over which exploration takes place. The data that gets stored and transmitted is inevitably the result of a person at some point making design decisions and engineering features about how that data will be analyzed.

3. THE INFORMATION ECOSYSTEM ATTACK SURFACE

Addresses Question 2: Preserving information integrity and mitigating the effects of information manipulation.

In cybersecurity research, it is common to describe risks in terms of an *attack surface*. An attack surface is, "a set of ways in which an adversary can enter the system and potentially cause damage" [17]. It provides a common vocabulary for describing a complex environment with respect to risk. Thinking about information integrity as an attack surface can accomplish the same thing. Having a common vocabulary can help guide R&D investments to address problems with information manipulation.

In the following sections, we briefly describe the nature of the attack surface for each component of the information ecosystem and highlight examples of R&D, including R&D at Sandia, that addresses the attack surface. It is not possible to comprehensively cover the research done at Sandia, much less the research done across the entire community. The examples are provided as illustrations of pressing R&D needs and are not comprehensive.

3.1. Decision

The Decision category of the information ecosystem is entirely concerned with the meaning of information in its context, used for some particular use case, regardless of its form. The attack surface of the Decision category deals with presenting information in such a way that when someone consumes and uses the information, they are unable to solve the problem for which they accessed the information in the first place.

One risk is related to data context and underlying assumptions. Data collection is done to support the needs of individuals and organizations and, as such, individuals must be able to make sense of the data in order to effectively utilize it. There are inherent data characteristics/attributes associated with the context under which the data was collected that ultimately impacts end users' trust and understanding of the data and lead to an association with the data's integrity. Borrowing from work in military intelligence and research in data fusion, four attributes are critical data characteristics that must remain associated with the data to support human understanding and the development of trust in the data: timeliness (time the data was collected), confidence (certainty of the data source), source (supplier or entity from which the information came), and accuracy (precision of the data) [5, 8]. These attributes must remain linked with the data during the collection process and during any associated data manipulations (such as data fusion and modeling) in order to maintain data integrity, as they influence how much trust is inherent in the data and how much weight may be assigned to conclusions derived from their output [9].

Disinformation is a component of the attack surface of information integrity. Disinformation is information shared with the intention of leading a consumer to a false conclusion. The simplest form of disinformation is a simple lie. Other forms of disinformation might include true statements that are misleading. There is significant interest in understanding the creation and spread of disinformation and its effect on our society.

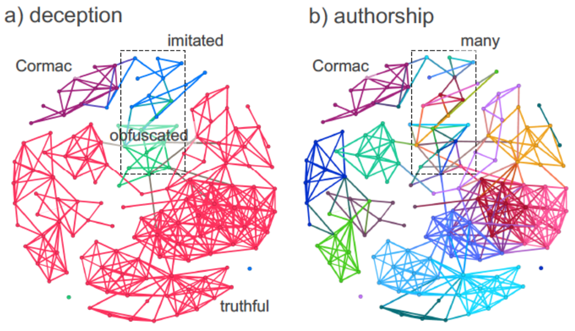


Figure 3-1. Clustering intentional deception (a) and by author (b) [23]

At Sandia, there is a growing body of research on disinformation. This includes research on deception detection within text. For example, an author attempting to pretend to be someone else by imitating their style. Figure 3-1 depicts previously published results [23] that show compression-based techniques over text can cluster by author. The use of these techniques for authorship detection is consistent with other previously published work [22]. Further, in the same work we conclude that the same techniques can be used for deception detection by means of clustering intentionally deceptive documents together. In more recent work, we have integrated psychological research with advances in compression-based analytics [3] to find disinformation in various genres of text [4].

3.2. Analysis

As mentioned in the introduction, Analysis is concerned with finding structure in data. The analysis pipeline takes information as input, performs a function like labeling or pattern detection on that information, and returns this augmented information as output. The output may also include summary information about the analysis. Problems arise when information is augmented or summarized in a way that compromises its integrity.

One technique commonly used for analysis is machine learning, which has a number of known security concerns. Attacking these weaknesses is called *Adversarial Machine Learning*. There are two categories of exploits of particular relevance to information integrity. Both categories target the input to an ML model, causing compromise in the output.

The first is evasion. Evasion attacks seek to avoid proper classification of an input item by slightly altering it. These are generally called *adversarial examples* [25]. A canonical example of an evasion attack is described in [11] where an alteration is made to an image that is not noticeable to humans, but changes the classification of the image by the machine learning model.

The second category is subversion. Subversion attacks seek to avoid proper classification of an input item by altering the model while it is being created. In the literature, many of these attacks use "data poisoning" where placing additional features on input images with different training labels can result in misclassification [25]. A model compromised in this way will perform exceptionally on a user's training and validation data, but poorly on specially selected inputs. One classic example of this type of subversion attack is described in [12]. Other examples of subversion attacks can be found in [15].

Adversarial Machine Learning and the defense against it ("Counter Adversarial Machine Learning") [25] highlight the classic security paradigm in which an attacker must only find one attack to be successful, while a defender must defend against all possible attacks to be successful. It follows that the attack surface of the Analysis category is very large. In particular, as solutions evolve to combat potential threats [26], new attacks are also developed [6] [24]. Ultimately continuing to build up defensive techniques to combat information manipulation in this category is only a partial solution. Holistic solutions that consider the full information ecosystem must instead be considered.

Analysis can be performed on a large amount of information so one output of the process is often a summary of the results. For example, performance statistics like accuracy, precision, recall, and F_1 score are ways to understand the output of an analysis at a glance. However, if inappropriate statistics are used for a given dataset or research question, the integrity of the summary information is lost. For instance, calculating accuracy on an imbalanced dataset may make an analysis appear more conclusive than it is. Consider the case when 90% of available information is irrelevant and 10% of information is relevant. If a model labels every piece of information as irrelevant, then the accuracy of that model is 90% even though it misses every instance of relevant information. This form of information manipulation is part of the attack surface of the Analysis category.

Consequences of the loss of information integrity at the Analysis category are observable in the Decision category of the information ecosystem. Examples include: diverted attention away from relevant pieces of information, wasted resources looking at irrelevant pieces of information, and inaccurate conclusions from manipulated data. However, the Analysis category of the ecosystem can be used to prevent some of these pitfalls. We suggest explainability as a possible defense against information manipulation.

One way to understand an analysis is by providing an explanation of where new information came from. For example, machine learning techniques that are "interpretable by design" are capable of determining which features led to a classification [1, 2]. Even when it is not possible to directly interpret a model, it is still possible to provide indirect explanations for predictions (even for models fit on functional data) [10]. These techniques give additional ways to assess the performance of the model and inspect outliers. While additional work is needed to ensure model explanations are accurate and indeed useful to decision makers (many current methods are shown not to be), there are proposed ways to address these concerns. Such methods include machine learning explainability techniques that take into account non-linear model interactions and design principles that take into account the usefulness of explanations to the end user [21]. This suggests, with some advancements, that explainability could be an effective tool for decision-makers to use to combat information manipulated in the Analysis category.

3.3. Data

The Data category of the information ecosystem is entirely concerned with how bits and bytes are moved and stored, regardless of the meaning of the information presented. The attack surface to information integrity represented by this category generally stems from the fact that the internet was built to be robust and fast, not secure. It is not the case that security was an afterthought. Early descriptions of internet protocol stacks mention the needs for security [7]. However, security was described as something that the internet model *supported* rather than *inherently contained*. Significant research has gone into developing security technologies for the data category and it is continuing to evolve. As more tools become cloud/service based it becomes more difficult to maintain data security without eroding capabilities in the Analysis and Decision domains.

One of the active areas of research has to do with the integration of cybersecurity techniques with experts in order to support the evaluation of large, constantly streaming data. Some of these efforts include developing platforms for testing and training on new techniques [18]. Sandia has also developed an Emulytics platform that simulates larger facilities to understand how to defend them [14].

4. CONCLUSION

Our information integrity ecosystem contains a significant attack surface that must be addressed to foster an environment of trust and resilience. Without this, individuals experience a compromised ability to think critically about information they encounter and consume. Many perspectives are required, including political, policy, social, and scientific, to tackle this problem. Interdisciplinary scientific R&D across the decision-analysis-data model is needed to effectively address how bytes are being protected and analyzed to support robust decision-making. Such considerations will enable us to not only understand the but also improve protections to ensure information integrity across the information ecosystem.

REFERENCES

[1] Sapan Agarwal and Corey M. Hudson. Probability series expansion classifier that is interpretable by design, 2017.

[2] Sapan Agarwal and USDOE. AWE-ML: Averaged weights for explainable machine learning v. 1.0, version v. 1.0, March 2019.

[3] Travis Bauer. Ngramppm: Compression analytics without compression. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2021.

[4] Travis Bauer, Lisa Gribble, and Nicole Murchison. Human constrained machine learning for deception detection in text. In *INFORMS Annual Meeting*, 2021.

[5] Erik P Blasch, Mike Pribilski, Bryan Daughtery, Brian Roscoe, and Josh Gunsett. Fusion metrics for dynamic situation analysis. In *Signal Processing, Sensor Fusion, and Target Recognition XIII*, volume 5429, pages 428–438. International Society for Optics and Photonics, 2004.

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

[7] Vinton G Cerf and Edward Cain. The DoD internet architecture model. *Computer Networks (1976)*, 7(5):307–318, 1983.

[8] Mica R Endsley, Betty Bolté, and Debra G Jones. *Designing for situation awareness: An approach to user-centered design*. CRC press, 2003.

[9] Eliezer Geisler, Paul Prabhaker, and Madhavan Nayar. Information integrity: an emerging field and the state of knowledge. In *PICMET'03: Portland International Conference on Management of Engineering and Technology Technology Management for Reshaping the World, 2003.*, pages 217–221. IEEE, 2003.

[10] Katherine Goode, Daniel Ries, and Joshua Zollweg. Explaining neural network predictions for functional data using principal component analysis and feature importance. *arXiv preprint arXiv:2010.12063*, 2020.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[13] Kelsey Harley and Rodney Cooper. Information integrity: Are we there yet? *ACM Computing Surveys (CSUR)*, 54(2):1–35, 2021.

[14] Corey Hudson. Emulytics in genome security: Use cases. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2021.

[15] Philip Kegelmeyer, M Timothy Shead, Jonathan Crussell, Katie Rodhouse, Dave Robinson, Curtis Johnson, Dave Zage, Warren Davis, Jeremy Wendt, "J.D." Justin Doak, Tiawna Cayton, Richard Colbaugh, Kristin Glass, Brian Jones, and Jeff Shelburg. Counter adversarial data analytics. Technical report SAND2015-3711, Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550, 5 2015.

[16] Butler W Lampson. Protection. *Proc. Fifth Princeton Symposium on Information Sciences and Systems*, pages 18–24, 1971.

[17] Pratyusa K Manadhata and Jeannette M Wing. An attack surface metric. *IEEE Transactions on Software Engineering*, 37(3):371–386, 2010.

[18] Kevin S Nauer, Seanmichael Yurko Galvin, and Tommie G Kuykendall. Tracerfire. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2016.

[19] Networking and Information Technology Research and Development (NITRD), National Coordination Office (NCO), and National Science Foundation (NSF). Request for information on federal priorities for information integrity research and development. *Federal Register*, 87(52), March 17 2022.

[20] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[21] Michael R. Smith, Erin Acquesta, Arlo Ames, Alycia Carey, Christopher Cuellar, Richard Field, and Trevor Maxfield. SAGE intrusion detection system: Sensitivity analysis guided explainability for machine learning. Technical Report SAND2021-11358, Sandia National Laboratories, September 2021.

[22] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[23] Christina L Ting, Andrew N Fisher, and Travis L Bauer. Compression-based algorithms for deception detection. In *International Conference on Social Informatics*, pages 257–276. Springer, 2017.

[24] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

[25] Jeremy D. Wendt. Position paper: Counter-adversarial machine learning is a critical concern. Technical report SAND2022-1493C, Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550, 2 2022.

[26] Xinqiao Zhang, Huili Chen, and Farinaz Koushanfar. TAD: Trigger approximation based black-box trojan detection for ai. *arXiv preprint arXiv:2102.01815*, 2021.

[27] Lina Zhou and Simon Lutterbie. Deception across cultures: Bottom-up and top-down approaches. In *International Conference on Intelligence and Security Informatics*, pages 465–470. Springer, 2005.

[28] Lina Zhou and Yu-wei Sung. Cues to deception in online Chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 146–146. IEEE, 2008.

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

SIFT, LLC

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.

RESPONSE TO: REQUEST FOR INFORMATION ON FEDERAL PRIORITIES FOR INFORMATION INTEGRITY RESEARCH AND DEVELOPMENT

SIFT, LLC

Misinformation and disinformation are longstanding problems that have recently grown in complexity due to enabling technologies. The lines of transmission are more numerous, rapid, and dynamic than ever. We propose technology and research questions aimed at empowering the characterization of information reliability, the production of more reliable information, as well as permeating the public square with new literacy, expectations, and standards related to information integrity. The central conceptual pillar of our approach is information provenance – structured data about how the information came to be. As one concrete application of provenance-based analytics, SIFT has developed Project7, an experimental, collaborative human-machine information analysis workspace. Project7 demonstrates that provenance graph representations and interactive provenance-based displays can help assess the origins, bias, confidence, and integrity of information. Its development has been informed by standards, directives, and metrics for integrity and rigor originating from the intelligence community. We propose to utilize information analysis workspaces such as Project7 as platforms for investigating information integrity research questions spanning intelligence analysis, journalism, and other content creation domains. Research in this area has implications for characterizing and increasing information integrity and improving consumers' resilience to misinformation and disinformation.

Introduction

Misinformation and disinformation are problems older than the printing press, but in the digital age they have swelled into massive challenges that hinder our ability to establish a commonly held set of basic facts with which to collectively debate and reason. The explosion of information pathways has rendered our previous defense mechanisms, which centered on fostering analytical rigor and integrity in the institutions with the most reach, insufficient - yet more important than ever. In this response we propose promising leads and timely questions that aim to raise the bar of informational integrity to a new high for all information producers, brokers, and consumers, by empowering them with tools and standards for analytical rigor.

Two critical efforts for mitigating the effects of misinformation and disinformation are (1) advancing the trustworthiness of information, and (2) enhancing the *informational immune systems* (i.e., their ability to discern and resist mis/disinformation) of individuals who consume information, drive commercial incentives for information producers, disseminate information through social networks, and increasingly participate in the production of journalism. These two pursuits are closely related, and we briefly discuss some connections between them before describing technological advances.

Trustworthiness & Perceived Credibility

For information consumers, a key asset is the availability of sources that are both *perceived* as trustworthy and are also *worthy* of that trust. Guillory and Geraci (2013) found that incorrect initial inferences, while powerful and persistent, can be corrected if the correction comes from a trustworthy source. In that study, trustworthiness (i.e, the perceived willingness to be accurate) was found to be more persuasive than expertise (i.e., the perceived ability to be accurate), which agrees with previous studies (McGinnies & Ward 1980; Lui & Standing 1989). Perceived credibility – closely related to but not synonymous with trustworthiness – has also been found to be associated with a higher likelihood of future engagement with a news source (Peifer & Meisinger 2021), which suggests that establishing credibility when disseminating information can have lasting benefits on impact that stretch beyond that piece of information alone. This evidence converges with the positive link between trust in a news source and loyalty to it, found by Nelson & Kim (2021). All else being equal, more trusted information brokers have greater potential to become resources for vetting, contextualizing and ultimately neutralizing misleading information. Broad-based trust is powerful, but rare because it is so arduous to build and so susceptible to self-destruction.

Improving Credibility, Reliability, and Transparency of Information Producers

One intuitive way to improve the credibility of information producers is to help them be more reliable (i.e., more accurate more often) by creating tools that enable them to conduct analyses with rigor (i.e., thorough, clear, and able to be validated and critiqued for confidence) (Zelik et al., 2010). Here we see a valuable role for software assistants that help track and publish *provenance* – data about the information's synthesis and the sources from which it came – and assist with critical-thinking tasks using that provenance. We illustrate this in later sections.

It would be ideal for these tools to also allow information producers (e.g., reporters, intelligence analysts, and their software analytics) to be able to "show their work" without compromising sensitive sources and methods, to facilitate better collaboration and transparency with consumers. For two decades, since the burgeoning of online news, blogging and the adage "transparency is the new objectivity," much research has gone into investigating the effect of transparency on credibility and trust in journalism. Some results have confirmed that there exists a positive relationship (Curry & Stroud 2019; Peifer & Meisinger 2021), but others suggest a more nuanced relationship. Karlsson and Clerwall (2018) found hyperlinks to be a form of transparency that was met with particularly positive response from readers because "hyperlinks make it possible to track down original sources and documents," among other reasons, while providing "negative user commentary" as transparency was counterproductive. Tandoc and Thomas (2017) found a negative effect of transparency on credibility when the form of transparency was the disclosure of biographical information about the author. Karlsson (2020) found that "participatory transparency" – the involvement of users in various stages of the reporting process, e.g. by sending pictures of events –increased source trustworthiness with the group of readers that had the most skeptical attitudes towards news media (low educated males), while it decreased trustworthiness with the least skeptical group (highly educated females). Transparency's real-world implementation is also a complicating matter. Despite transparency's efficacious origins in the blogging world (Lasika 2004), Koliska and Chadha (2016) later studied newsrooms and found that the practice of transparency in corporate journalism had become largely performative, having been mandated and implemented without the direct involvement of journalists. A common consensus of the research into transparency's effect on credibility seems to be that "transparency" encompasses too many elements for sweeping conclusions to be made, and that different audiences favor different forms of transparency.

The expectations and effectiveness of different transparency and disclosure strategies also may be highly dependent on the underlying analysis and reporting processes. This is consistent with the finding of Diakopoulos and Koliska (2017) that information consumers desire certain types of transparency from algorithmic (i.e., machine-generated) reporting, including information about the data, the model, inference methods, and the availability of a public-facing interface into these aspects.

The previous work described above suggests that trustworthiness of information – and consumers' ability to characterize it – can be enhanced by tools that support (1) more rigorous, collaborative analysis – a form of quality assurance – and (2) customizable transparency about the genesis of the information, including machine reasoning. These are some of the primary motivations behind SIFT's ongoing R&D on provenance and the Project7 human-machine analysis workspace. Below we discuss these technologies, their relevance to intelligence analysis and journalism, and relevant standards and metrics for analytical rigor. We outline some connections between these tools and ideas for enhancing the public's immunity to manipulated information, identifying relevant research questions.

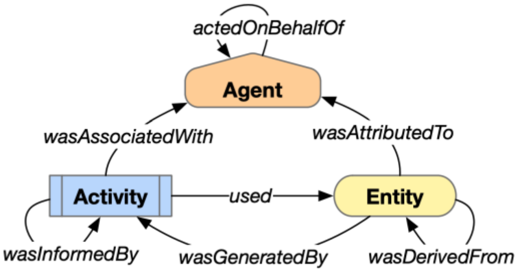


Figure 2: The PROV data model.

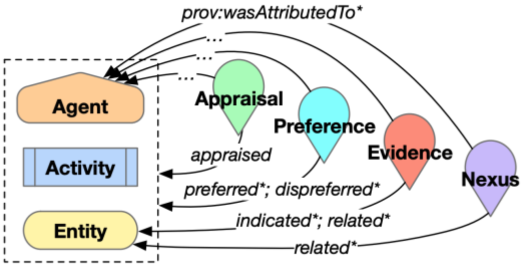


Figure 1: The DIVE ontology.

Provenance

The provenance of a piece of information refers to data about its origins. Formal provenance ontologies consist of the types of – and types of relationships between – elements that can play a role in these origins.

SIFT's provenance research products are built around the W3C-recommended PROV data model (Figure 1), which contains three types of elements: entities, activities, and agents. *Entities* are fixed real or hypothetical things, such as records, assertions, databases, etc., which can be input to or outputs of activities. *Activities* are processes that occur over a period of time, such as inference actions or other procedures performed by human or machine. *Agents* are those actors that perform activities, whether they be humans, organizations, web services or machine learning modules.

The predicates in the PROV data model relate particular types of elements to other types of elements. PROV can represent that an entity **was derived from** another entity, **was attributed**

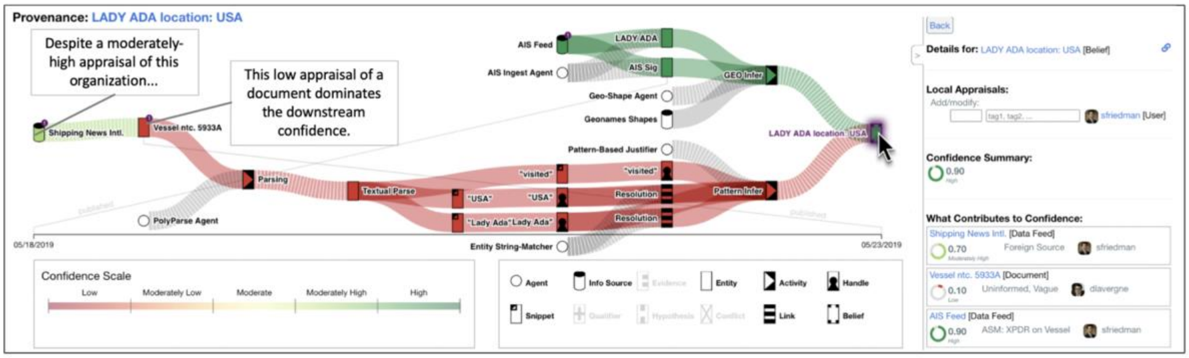
to an agent, or **was generated by** an activity. It can represent that an activity **used** an entity, **was informed by** another activity, or **was associated with** an agent. Finally, it can represent that an agent **acted on behalf of** another agent. All of these relationships help represent directed information dependency, e.g., an entity that **was generated by** an activity can be seen as immediately *downstream* from that activity, and that activity would in turn be immediately downstream of any entities that it **used**. In this way, the PROV data model can be used to represent the full inferential provenance for a piece of data (e.g., an assertion) as a dependency graph, called a provenance graph.

Provenance alone helps express the structure of an analysis, but it does not express the information integrity considerations of information diversity, confidence, alternatives, assumptions, gaps, conflicts, likelihood, or biases. To help express these considerations, SIFT and BBN extended the PROV data model with the DIVE ontology (Friedman et al. 2020; Figure 2). DIVE adds four classes of provenance elements, all of which represent judgments that are attributed to agents (**wasAttributedTo**). An *Appraisal* is a confidence judgment about any other (**appraised**) element, with attributes for confidence, likelihood, bias, and reliability. A *Preference* is a judgment about the relative quality between one (**preferred)** element and another (**dispreferred**) element, all else being equal. *Evidence* is a judgment about the diagnosticity of one (**related**) entity on another (**indicated**) entity, e.g., evidence toward a hypothesis. A *Nexus* is a judgment about the mutual coherence or conflict within a set of (**related**) entities (i.e., that the entities in set have high or low joint likelihood). These extensions capture local quality judgments within the provenance graph that can support global information integrity assessments. Follow-on research should extend DIVE to cover a fuller range of human explanation and argumentation, in such a way that fits into a dependency graph.

The provenance graph can serve as a substrate for critical thinking about sensitivity, confidence, information necessity and sufficiency, and impact, using ATMS-inspired algorithms (Forbus & de Kleer, 1993; Friedman et al. 2021). For example, new assessments about the reliability of sources upstream can potentially justify shifts in judgments about the quality of resulting inferences downstream. Similarly, semantic information tags such as INTs (*OSINT***,** *HUMINT***,** *IMINT***,** etc.) source types (e.g., *Online News*, *Social Media*), operation types (e.g., *NLP*, *Named Entity Recognition***,** *Machine Learning*), and operating assumptions (e.g., inferring a vessel's location via a transponder signal implies the transponder is *on* the vessel) may be added to individual nodes, and algorithms can propagate these downstream. These propagation algorithms allow the user to (temporarily) remove a source to assess its downstream impact on their conclusions, propagate confidence downstream from sources to estimate the confidence of conclusions, and assess the diversity, gaps, and assumptions in downstream conclusions.

Provenance Graphs in Project7

Project7 is an experimental human-machine information analysis workspace supported by multiple efforts for high-integrity information analysis (Friedman et al. 2021). Project7 allows multiple users to collaborate on the generation and validation of competing hypotheses, with an interactive view of the provenance graph playing a central role.


Provenance: LADY ADA location: USA

Despite a moderately-high appraisal of this organization...

This low appraisal of a document dominates the downstream confidence.

Back

Details for: LADY ADA location: USA [Belief]

Local Appraisals:

Add/modify:

tag1, tag2, ... sfriedman [User]

Confidence Summary:

0.90 High

What Contributes to Confidence:

Shipping News Intl. [Data Feed]
0.70 Moderately High Foreign Source sfriedman

Vessel ntc. 5933A [Document]
0.10 Low Uninformed, Vague dlavergne

AIS Feed [Data Feed]
0.90 High ASM: XPDR on Vessel sfriedman

Figure 3: A provenance graph in Project7 for the inference that the ship LADY ADA was located in the USA.

Figure 3 shows an example provenance display for the hypothesis that the fictional vessel *Lady Ada* is located in *USA*. The hypothesis (Figure 3, right) was generated by two alternative inferential paths through the provenance, one with a high confidence and the other with low confidence, resulting from user appraisals upstream. Project7 allows users to explore alternative, experimental approaches to propagating quality judgments through the graph. In Figure 3, confidence propagates rightward through junctions as if they are an AND (min input value) or an OR (max input value) depending on whether the incoming relations are semantically necessary vs. sufficient. The interface allows users to conduct sensitivity analysis by excluding elements by their identity or properties, including users. In this example, excluding the user *sfriedman* would invalidate their appraisals for *Shipping News Intl.* and the *AIS Feed*, and the high confidence would cease to flow through the upper path, ultimately dropping the confidence in the hypothesis to low (via only the bottom path). This helps users quickly answer *"what if?"* questions and investigate the global impact of particular elements in the analysis.

Using provenance as an interactive substrate facilitates user-driven exploration of the sensitivity of inferences to hypothetical changes in their related assumptions, processes, data sources, and key contributors. Expanding this approach and applying it more broadly has the potential to revolutionize how content is collaboratively generated and disseminated with a suitable level of transparency. An open research question is how the machine can proactively contribute to this exploration, acting as a critical-thinking assistant and radically improving rigor by mitigating cognitive biases with low cost, high coverage, and high throughput.

Intelligence and Journalism

Intelligence analysis and journalism share many of the same concerns and best practices. Both enterprises collect information, and, where possible, synthesize meaningful inferences for decision makers in the information ecosystem. Both have access to specialized sources and collection methods with diverse capabilities and technical risks. Both have reasons to "show their work" (assumptions, sourcing, argumentation) as much as possible while also protecting sources and methods. Both practices can be collaborative, which adds communication risks to the reporting process, and creates an imperative to keep fulsome and explicit records.

A significant difference is that journalists' output, when ready, is typically dispersed broadly into the public information sphere, whereas intelligence analysts' outputs have a more variegated audience and distribution methods governed by security controls. The provenance graph is itself a useful data structure for propagating these controls from sources and methods to assessments. In this dimension, we see the practice of journalism as subsumed by the practice of intelligence analysis, and thus by starting as an intelligence analysis workspace, ideas from Project7 are well positioned to generalize to journalism. In both domains, information integrity tools can help maintain standards for rigor in an environment where human operators are susceptible to intrinsic cognitive biases, incentive structures, and attentional limitations.

We posit that a long-term goal for building resiliency to misinformation and disinformation should be to enhance the public's literacy, expectations and tools for provenance and rigor in the information they consume. This shift can be driven in part by information producers, with the help of our research and development.

Supporting Rigorous Analysis

Provenance-based analytics can advance the trustworthiness of information by supporting aspects of rigor for analysts in the intelligence community and journalism. Here we draw attention to noteworthy properties of analytic rigor that motivate SIFT's provenance-guided analytics, and to others where analytics like Project7 are not yet sufficient.

The Office of the Director of National Intelligence (2021) has published intelligence community directives (ICDs). ICD 203 contains nine tradecraft standards intended as guides for achieving analytic rigor and quality. Similarly, Zelik et al.'s (2010) eight attributes of analytical rigor also provide guidance for high-quality information analysis. We review these standards and metrics of rigor to illustrate how provenance addresses integrity and rigor in this domain.

ICD 203's standard two advises that intelligence "Properly expresses and explains uncertainties associated with major analytic judgments." As discussed above, provenance-based analytics allows users to appraise elements in the graph with a confidence scale that has quantitative and qualitative aspects, and these confidences propagate through the graph using a set of alternative propagation schemes.

Standard three advises that an analytic product "Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments." As shown in Figure 3, Project7 uses iconography to display canonical graph nodes (e.g., beliefs) and semantic tags, including underlying assumptions, to express source diversity and facilitate visual filtering.

Standard four, and Zelik et al.'s "Hypothesis exploration" metric, concern the analysis of alternative explanations and possibilities. Here provenance is only marginally useful, since it can support comparative reasoning about *existing* alternatives but cannot support the automatic *generation* of alternative hypotheses. This raises a high-impact research question: how to automatically generate alternative hypotheses that plausibly explain the data? Case-based reasoning (CBR) over provenance graphs may address this research challenge, but this would not address the "cold-start" version of the problem, which likely calls for domain-specific reasoning and/or common-sense knowledge.

According to standard six: "Products should state assumptions explicitly when they serve as the linchpin of an argument or when they bridge key information gaps. Products should explain the implications for judgments if assumptions prove to be incorrect. Products also should, as appropriate, identify indicators that, if detected, would alter judgments." This is another aspect of rigor where provenance has a good start, allowing users to interactively discover the downstream effects of assumptions proving incorrect. It is poised to take a further step in this direction – automatically searching for high-impact, linchpin assumptions (and evidentiary linchpins in general) – to present an explicit ordered list to users. The last element – identifying indicators that would alter judgments were they detected – is a natural extension of Project7 for indicators that are already represented in the provenance, but difficult for hypothetical (missing) indicators, short of perhaps incorporating CBR to use previous cases to hypothesize relevant indicators. Related to this standard, Zelik et al.'s "Sensitivity Analysis" attribute of rigor calls for evaluating the strength of an assessment given variations in source reliability and uncertainty. Project7 already supports human-led sensitivity analysis and is well-positioned to automate the search for sensitivities and quantify them. A relevant research question is: what are useful measures of sensitivity, in general or with respect to certain types of data? A sensitivity measure might have to identify sensitivities that an expert analyst or journalist would find relevant, and also express the sensitivity to the lay public, to improve data integrity literacy (more below).

Standard eight is noteworthy: "Analytic products should apply expertise and logic to make the most accurate judgments and assessments possible." Provenance does not vet the logic of an argument, nor can it verify that relevant expertise has been captured in the argument. A provenance graph instead provides a record of argumentation, which can act as scaffolding for human users to know where to investigate for logical soundness. This seems out of reach for the machine without a much richer ontology and more pervasive reference-resolution capabilities.

Standard nine is about incorporating visual information where appropriate. Project7 adheres to this standard by providing an interactive view of the provenance graph, as well as by linking items in the graph to other available views, e.g., imagery, an interactive document view for text/NLP, a map view for items with a geographical aspects, a timeline for temporally situated elements, and more.

Zelik et al.'s "Information Validation" metric is about actively and systematically vetting collected data with multiple independent, credible sources, and seeking data with convergent evidence. Project7 is able to simultaneously search multiple structured data stores (e.g., Wikidata, DBpedia, Open Street Maps), and is equipped with NLP tools that help synthesize their results, e.g. named entity recognition. It offers a canvas for constructing a concept web that integrates pieces of gathered evidence. These features can help draw out a convergence of evidence where it exists, but their coordination must be performed by a human operator. More sophisticated reasoning may enable a more proactive role for the machine-as-data-validator.

Also of note is Zelik et al.'s "Explanation Critiquing" metric of analytical rigor, which has to do with seeking feedback on an entire analysis. One positive indicator is the use of "devil's advocacy" to challenge hypotheses and explanations. A multi-user environment with support for differing appraisals like Project7 is a useful tool in this regard. Better would be automated or semi-automated devil's advocacy, which points to an interesting research question: Given a provenance graph, how can rigorous counter arguments be automatically gathered from the public sphere?

Building Public Resistance: Expectations, Literacy, and Standards

We believe an essential pillar of resiliency to information manipulation is building up widespread immunity among the public. This endeavor involves establishing higher expectations for trustworthy information, new forms of literacy, and standards and practices that have the potential to gain traction. Below we outline ideas that connect these directions to our discussion of provenance and rigor above.

First, there is a potential link between advancing the trustworthiness of information by supporting analytical rigor and raising the public's expectations about information integrity. Empowering media organizations, intelligence analysts, and other information producers to output more reliable information may gradually build public trust in the institutions that exercise this power. In short, by supporting rigorous reporting we can raise the bar for information integrity. This may in turn make information consumers less forgiving of outlets that produce less reliable or manipulated information. Another way to drive expectations for information integrity would be to make available a public-facing interface for analyzing provenance. Imagine if authors – equipped with a provenance management interface – were able to export the full provenance for their reporting either as metadata embedded in the article or as a linked data file. The savvy information consumer might then load the article into the online provenance interface of their choice, allowing them to probe deep analytical questions and measure the information for rigor, ultimately increasing trust in information that is worth trusting.

A related pursuit is fostering public literacy about information integrity. In previous sections, we have touched on a relatively simple provenance ontology, quantitative measures of uncertainty (confidence and likelihood), and concepts from analytic rigor such as linchpin assumptions, judgment sensitivity, and hypothesis exploration. All of these would be useful to inject into the common parlance of information exchange, and the simplest way to do so is from the supply side. To start to build expectations, analysts and reporters could provide meta information to summarize the provenance, certainty, and analytical process behind the assessments that they produce. Provenance-based analytics can help by facilitating or even automating this summarization. A related subproblem is the generation of natural language descriptions of provenance graphs, which was achieved by the PROVglish architecture (Richardson & Moreau 2016). Another related research question would be how to control and automatically tailor the level of detail in the summarization. For the intelligence community this concern relates to clearances, "tear lines", and (for journalists as well) protecting sources, but it is also a relevant concern for improving public literacy about information integrity, since the sharing context might inform what level of detail is palatable for the consumer. For example, a tweet might call for a more brief summary than a feature article. Different topics may also call for different forms of sharing.

It would also be helpful to establish standards to guide expectations and ground this new literacy. A common system of interpretable tags to characterize provenance structure, certainty, assumptions, sensitivities, alternatives, etc., would make our new information-integrity literacy easier to convey, easier to use for quick comparisons and sorting, more salient, and more trendy. It would also be worth trying to establish a standard set of metrics to characterize information integrity, e.g. a measure of an assessment's sensitivity as described in the previous section. It is important that these information integrity tags and metrics are as simple as possible while capturing a useful level of detail, to give them the best chance of widespread use.

Finally, provenance technologies could improve the evolving practice of involving the public in the production of journalism, by providing structure and managing risk. Crowdsourced journalism, participatory journalism, citizen journalism and grassroots journalism have intersecting definitions, but their increasing relevance makes clear that individuals – even those outside of journalistic institutions – have more opportunities to directly participate in the production of information that is received with higher legitimacy. These practices are not only vehicles for public literacy, but pathways by which literacy feeds back into the trustworthiness of information, making them a compelling topic of research. Crowdsourced investigations have proven powerful and unwieldy, marked by life-saving successes as well as unjust cases of misidentification and vigilantism (Venkatagiri, 2021). We can see some potential benefits of crowdsourced journalism in metrics of analytic rigor discussed above – namely hypothesis exploration, information validation, and explanation critique. Indeed, Aitamurto (2019) identified ways in which crowdsourced reporting can benefit the journalistic norms of accuracy, objectivity, and transparency. However, that paper also contained a discussion of its risks to the norms of accuracy, objectivity, and autonomy, in which a theme was the lack of structure, both in the process – which works better when led by the journalist – and in the crowdsourced information itself – which has the potential to overwhelm in unstructured form. Collaborative frameworks based on provenance ontologies could deliver the needed structure on both fronts and should be studied in this context.

Conclusions

We have targeted two interrelated questions that are critical for information integrity in today's ecosystem: (1) How can we advance the trustworthiness of information? (2) How can we boost the informational immune systems of the public? Provenance – data that describes the origins of information – is highly relevant to both questions. SIFT's formal treatment of provenance as a dependency graph has been useful for computing answers to questions of impact. We illustrated its utility by describing its role in Project7, an experimental intelligence analysis workspace that also has potential relevance to journalism and the public square. Highlighting the relevance of provenance-based analytics to tradecraft standards and rigor metrics elucidated where these tools are already innovative in advancing the trustworthiness of information and where they have potential to push it further. We finished by discussing potential benefits of these technologies for public resilience, namely higher expectations, better literacy, new standards, and collaborative frameworks for participatory journalism based on provenance.

References

Aitamurto, T. (2019). Crowdsourcing in journalism. In *Oxford Research Encyclopedia of Communication*.

Curry, A. L., & Stroud, N. J. (2021). The effects of journalistic transparency on credibility assessments and engagement intentions. *Journalism*, 22(4), 901-918.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital journalism*, 5(7), 809-828.

Forbus, K. D., & De Kleer, J. (1993). *Building problem solvers* (Vol. 1). MIT press.

Friedman, S., Rye, J., LaVergne, D., Thomsen, D., Allen, M., & Tunis, K. (2020). Provenance-based interpretation of multi-agent information analysis. *Proceedings of TaPP*.

Friedman, S. E., Rye, J., McLure, M., Wauck, H. C., Patel, P., Wheelock, R., Valovage, M., Johnston, S. & Miller, C. (2021, October). Provenance as a Substrate for Human Sensemaking and Explanation of Machine Collaborators. In *2021 IEEE International Conference on Systems, Man, and Cybernetics* (SMC) (pp. 1014-1019). IEEE.

Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition*, 2(4), 201-209.

Karlsson, M. (2020) Dispersing the Opacity of Transparency in Journalism on the Appeal of Different Forms of Transparency to the General Public, *Journalism Studies*, 21:13, 1795-1814, DOI: 10.1080/1461670X.2020.1790028

Karlsson, M., & Clerwall, C. (2018). Transparency to the Rescue? Evaluating citizens' views on transparency tools in journalism. *Journalism Studies*, 19(13), 1923-1933.

Koliska, M., & Chadha, K. (2016). Digitally outsourced: The limitations of computer-mediated transparency. *Journal of Media Ethics*, 31(1), 51-62.

Lasica, J. D. (2004). Transparency begets trust in the ever-expanding blogosphere. *Online Journalism Review*, 12.

Lui, L., & Standing, L. (1989). Communicator credibility: Trustworthiness defeats expertness. *Social Behavior & Personality: an international journal*, 17(2).

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, 6(3), 467-472.

Nelson, J. L., & Kim, S. J. (2021). Improve trust, increase loyalty? Analyzing the relationship between news credibility and consumption. *Journalism Practice*, 15(3), 348-365.

Office of the Director of National Intelligence. (2021). "Intelligence Community Directives," https://www.dni.gov/index.php/whatwe-do/ic-related-menus/ic-related-links/intelligence-communitydirectives, 2021.

Peifer, J. T., & Meisinger, J. (2021). The value of explaining the process: How journalistic transparency and perceptions of news media importance can (sometimes) foster message credibility and engagement intentions. *Journalism & Mass Communication* Quarterly, 98(3), 828-853.

Richardson, D. P., & Moreau, L. (2016). Towards the domain agnostic generation of natural language explanations from provenance graphs for casual users. In *International Provenance and Annotation Workshop* (pp. 95-106). Springer, Cham.

Tandoc Jr, E. C., & Thomas, R. J. (2017). Readers value objectivity over transparency. *Newspaper research journal*, 38(1), 32-45.

Venkatagiri, S., Gautam, A., & Luther, K. (2021). Crowdsolve: Managing tensions in an expert-led crowdsourced investigation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-30.

Zelik, D. J., Patterson, E. S., & Woods, D. D. (2010). Measuring attributes of rigor in information analysis. *Macrocognition metrics and scenarios: Design and evaluation for real-world teams*, 65-83.

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).

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

WITNESS

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.

May 15, 2022

Re: WITNESS RFI Response: Information Integrity R&D (Docket Number NSF - 2022-05683)

WITNESS takes this opportunity to respond to the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) and National Science Foundation (NSF) Request for Information on Federal Priorities for Information Integrity Research and Development.

About WITNESS

WITNESS is an international human rights organization that helps people use video and technology to protect and defend their rights. As a key element of that we ensure that vulnerable and marginalized communities are equipped with the skills, tools and infrastructure to create information with integrity, and challenge misinformation and disinformation. Working across five regions (Asia and the Pacific, Latin America and the Caribbean, the Middle East and North Africa, Sub-Saharan Africa, and the United States) alongside those most excluded or at-risk, our teams identify gaps, design solutions, provide guidance, and co-develop strategies. We then scale this work globally on a systems level, sharing what we learn with communities facing similar issues and advocating the needs of vulnerable and marginalized communities to technology companies and other influential stakeholders to ensure they are translated into policies, governance and solutions. As a critical part of that work over the past decade we have engaged in research, prototyping and action around critical issues in information integrity - including two foci: i) appropriate ways to develop content authenticity and provenance infrastructure and ii) counter emerging forms of visual deception such as malicious deepfakes via solutions grounded in a global, human rights framework.

More context on our work in this area can be found at: https://lab.witness.org/projects/synthetic-media-and-deep-fakes/ and https://lab.witness.org/ticks-or-it-didnt-happen/

WITNESS is also a member of the Coalition for Content Provenance and Authenticity (C2PA) where we support efforts to promote a human rights framework in the design of these specifications for content provenance and information integrity. WITNESS is the co-chair of the Threats and Harms Taskforce where it leads a harms, misuse and abuse assessment. This submission is not in our capacity as a C2PA member, however we reinforce aspects of the C2PA framework below.

1. Understanding the information ecosystem: *There are many components, interactions, incentives, social, psychological, physiological, and technological aspects, and other considerations that can be used to effectively characterize the information ecosystem. What are the key research challenges in providing a common foundation for understanding information manipulation within this complex information ecosystem?*

WITNESS supports research focused on two elements that are under-researched in the current environment.

The first is focused work on the prevalence and appropriate approaches to handling audiovisual information manipulation challenges. Existing research has a heavier focus on text, while video and audio continue to grow in prominence as vectors for information manipulation.

The second area is a focus on understanding how information manipulation affects vulnerable and marginalized communities in the US and globally and how interventions are assessed for their proportional or disproportionate impact on vulnerable communities. This area includes understanding how information integrity interventions that apply to US-based platforms – or are developed in US contexts – function in a diverse set of global geographies and contexts with different standards of rule-of-law, adherence to human rights and attention from social media platforms.

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. What are the key gaps in knowledge or capabilities that research should focus on, in order to advance these objectives? What* *are the gaps in knowledge regarding the differential impact of information manipulation and mitigations on different demographic groups?*

WITNESS recommends that the following be prioritized as a methodological approach to preserving information integrity and mitigating the effects of information manipulation amid a range of approaches that includes literacy, tools and architectures/standards for understanding manipulation as well as content integrity, authenticity and provenance.

1) Research that from early-on centers a human-rights based approach to understanding potential harms, misuses and unintended consequences of interventions, particularly as they impact vulnerable and marginalized populations. From our context as members of the C2PA, this coalition's efforts provide a good example of early work to identify potential harms that may come from these specifications, especially to those that may be marginalized and particularly vulnerable to information manipulation. Additional research on information interventions is critical that centers consultations with individuals, communities and institutions likely to be impacted by interventions, and to place emphasis on those who already face similar systemic harms. Broad, ongoing, multi-disciplinary consultations are a necessary basis to understand how information manipulation affects different demographic groups, and how standards and practices may avert harm and open up new opportunities to tackle mis- and disinformation and to raise the trustworthiness of authentic content.
2) As noted above, understanding of the global implications of information interventions that are unevenly applied.
3) Research into strategies for improving access to emergent solutions to information integrity issues that inherently require access restrictions (e.g. for cybersecurity and efficacy reasons) but also contain an approach to ensure that the most vulnerable have access to protections. An example is in the area of deepfakes detection where WITNESS has identified a looming problem (see https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/) for media and civil society globally, and where controlled access to tools must be balanced with global equity and need.

3. Information awareness and education: A key element of information integrity is to foster resilient and empowered individuals and institutions that can identify and abate manipulated information and create and utilize trustworthy information. What issues should research focus on to understand the barriers to greater public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways?

WITNESS has identified a more transparent information ecosystem as one of the frameworks that can increase public awareness about manipulated information and facilitate educational processes, provided that this transparency can be achieved in an opt-in manner that facilitates privacy and recognizes a diverse set of global contexts for how individuals choose to share information and the regulatory and legal contexts in which they share information.

Although transparency can be understood and expressed in a myriad of ways, WITNESS has been exploring three options: the disclosure of provenance information (such as the C2PA), the inclusion of forensic traces (particularly for new forms of synthetic media where photorealistic manipulation is increasingly hard to spot) and labeling of information as it circulates. Of those, we discuss the first one here.

The same questions mentioned above could be reformulated in the context of provenance information as a potential solution: How could provenance information help increase public awareness of information manipulation? What challenges should research focus on to support the development of effective educational pathways with provenance information?

To answer these questions, there is a need to continue researching how provenance tools and standards could be made accessible to as many people, communities and entities that may need it, all the while ensuring their privacy and other human rights.

There is also a need to develop prototypes that cater to different potential content creators in order to understand what would spark adoption, what would promote privacy-oriented designs and usages, and what are the emerging threats and harms. There is also the need to research how consumers process provenance indicators, and what are the expressions that promote trust and what are those that could undermine the trustworthiness of authentic content.

WITNESS provides further insights into research dilemmas and questions in our 2019 research report 'Ticks or It Didn't Happen: Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia' (https://lab.witness.org/ticks-or-it-didnt-happen/)

4. Barriers for research: Information integrity is a complex and multidisciplinary problem with many technical, social, and policy challenges that requires the sharing of expertise, data, and practices across the full spectrum of stakeholders, both domestically and internationally. What are the key barriers for conducting information integrity R&D? How could those barriers be remedied?

Inclusion more broadly in prioritization of potential threats and solutions for information integrity R&D is critical. WITNESS has pursued a deliberate multi-stakeholder process in our own work that focuses on centering a mix of lived, expert and technical expertise in relation to current and emerging information integrity issues to ensure that information integrity risks and challenges are understood in context, and solutions are prioritized within R&D based on a diverse set of inputs.

In the context of standards development bodies (SDOs), participation is restricted and they often lack the diversity of knowledge, experiences and identities that should inform the design of specifications that would have a determining role in any ecosystem. This is to a certain degree expected, since the work and discussions in these bodies tends to be of a highly technical nature. One remedy for this is to actively seek to 'translate' technical work into more accessible information aimed for broader audiences. As one example of proactive action in this regard, C2PA has made efforts to this end by publishing an 'Explainer' alongside the specifications.

Looking towards the future, more of these 'translations' are necessary, accompanied by committed efforts to socialize this information and to continuously engage with as many stakeholders as possible.

5. Transition to practice: How can the Federal government foster the rapid transfer of information integrity R&D insights and results into practice, for the timely benefit of stakeholders and society?

The Federal government can facilitate transfer of R&D insight into practice by active engagement with a diverse range of stakeholders designing and implementing interventions at a local level in the US as well as globally. The prioritization of vulnerable and marginalized stakeholders in this work is both important from a harm reduction perspective and ethical perspective, but also likely to lead to the most effective long-term impact.

6. Relevant activities: What other research and development strategies, plans, or activities, domestic or in other countries, including in multi-lateral organizations and within the private sector, should inform the U.S. Federal information integrity R&D strategic plan?

WITNESS notes the analysis in the recent UNESCO Broadband Commission report to which we contributed that highlights critical questions around information integrity and misinformation/disinformation questions from a freedom of expression human rights perspective: 'Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression' https://www.broadbandcommission.org/publication/balancing-act-countering-digital-disinformation/ and also notes the recent/upcoming work by the UN Special Rapporteur on Freedom of Expression in this area.

7. Support for technological advancement: How can the Federal information integrity R&D strategic plan support the White House Office of Science and Technology Policy's mission:

- Ensuring the United States leads the world in technologies that are critical to our economic prosperity and national security; and
- maintaining the core values behind America's scientific leadership, including openness, transparency, honesty, equity, fair competition, objectivity, and democratic values.

WITNESS has no comments on this area.