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

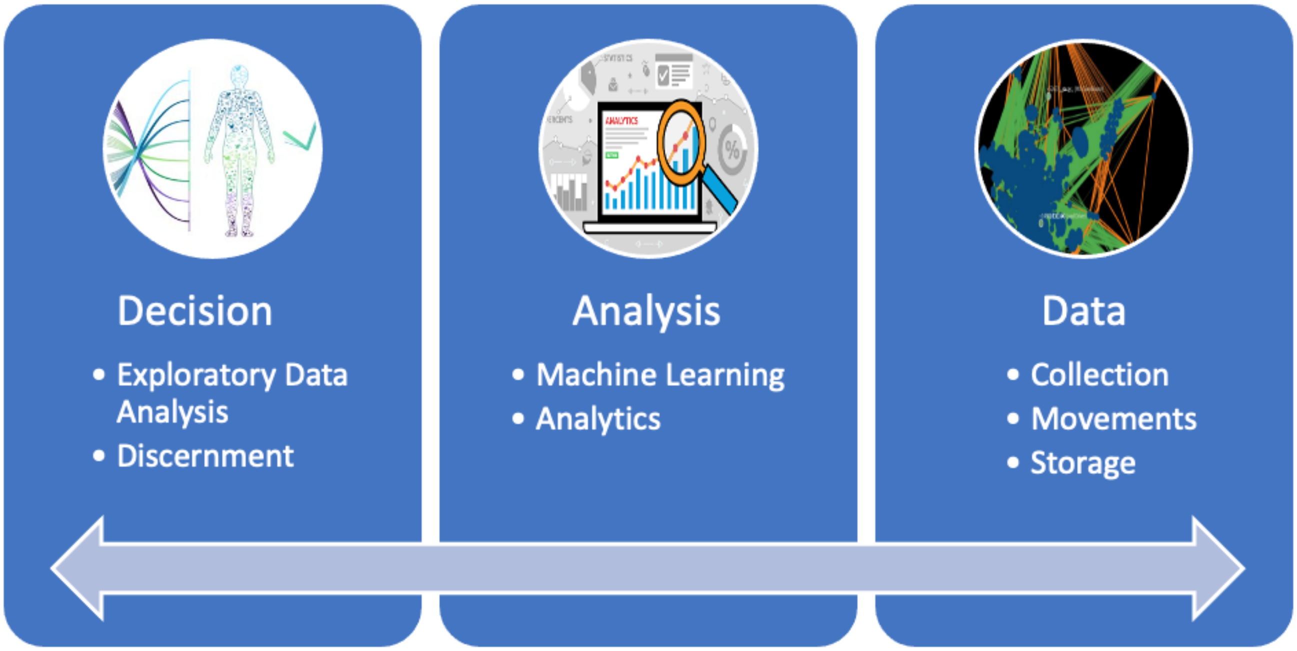

Figure 2-1. Model of the information ecosystem.

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.

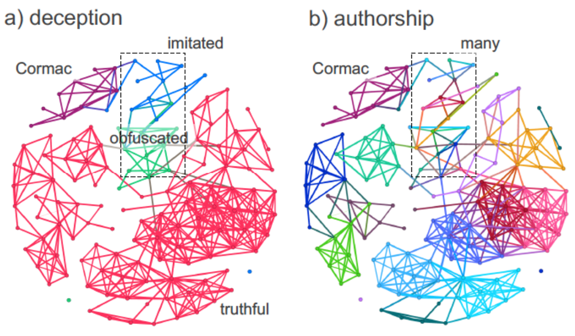

Figure 3-1. Clustering intentional deception (a) and by author (b) [23]

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.