ROADMAP FOR RESEARCHERS ON PRIORITIES RELATED TO INFORMATION INTEGRITY RESEARCH AND DEVELOPMENT

A Report by the

INFORMATION INTEGRITY RESEARCH & DEVELOPMENT
INTERAGENCY WORKING GROUP

NETWORKING & INFORMATION TECHNOLOGY RESEARCH &
DEVELOPMENT SUBCOMMITTEE

of the

NATIONAL SCIENCE & TECHNOLOGY COUNCIL

December 2022

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, health, foreign relations, and the environment, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal R&D in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure that science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About the Subcommittee on Networking & Information Technology Research & Development

The Networking and Information Technology Research and Development (NITRD) Program has been the Nation's primary source of federally funded work on pioneering information technologies (IT) in computing, networking, and software since it was first established as the High Performance Computing and Communications program following passage of the High Performance Computing Act of 1991. The NITRD Subcommittee of the NSTC guides the multiagency NITRD Program in its work to provide the R&D foundations for ensuring continued U.S. technological leadership and for meeting the Nation's needs for advanced IT. The National Coordination Office (NCO) supports the NITRD Subcommittee and its Interagency Working Groups (IWGs) (https://www.nitrd.gov/about/).

About the Information Integrity R&D Interagency Working Group

The Information Integrity R&D (IIRD) Interagency Working Group (IWG) provides a forum for interagency planning and coordination for R&D investments in information integrity, identifying research gaps, defining future research directions, and encouraging public-private partnerships in addressing information integrity challenges.

About This Document

This Roadmap for Researchers on Priorities Related to Information Integrity Research and Development was developed by the IIRD IWG of the NITRD Subcommittee. The Roadmap aims to coordinate and guide federally funded R&D in information integrity; it identifies six interrelated priorities for developing (1) models and measurements of information ecosystems, (2) safeguards that assist people, (3) technologies to enhance the integrity of information exchange, (4) strategies to address manipulated-information campaigns, (5) data access and partnerships, and (6) connection of research to policy and practice.

Copyright

Note: Any mention in the text of commercial, non-profit, academic partners, or their products, or references is for information only; it does not imply endorsement or recommendation by any U.S. Government agency.

Preface

Accurate and reliable information is central to the well-being of people and society. Manipulated information can cause harm across many social domains, including national security, local and national crisis response efforts, human rights and protections, and people's individual health and safety.

Recognizing these possible harms and the complexities of today's information environment, a broad group of Federal agencies, together with the Office of Science and Technology Policy, came together to develop the Roadmap for Researchers on Priorities Related to Information Integrity Research and Development. The goal of this Roadmap is to share Federal Government research priorities in the area of information integrity, focusing on expanding access to high-integrity information while minimizing harm.

Importantly, the Roadmap is not a policy directive. It does not make any recommendations regarding laws. It also does not include any requirements or regulations. Instead, its purpose is to stimulate research that can be used to strengthen the role of information ecosystems in the open exchange of ideas where healthy debate and free expression thrive.

We in the Federal Government developed this Roadmap through sustained discussion across Federal agencies and through numerous consultations with academic researchers, commercial entities, international partners, those adversely affected by corrupted information, former government employees across the political spectrum, and others seeking to address information integrity challenges.

It is important to acknowledge that protections for free speech and human rights are fundamental. Therefore, this Roadmap encourages research approaches that involve communities as active participants and that recognize responsibilities in the public and private sectors.

This document is a starting point for conversations with the research community and those who wish to contribute to research. We look forward to earnest, substantive, respectful dialogue about this Roadmap, about gaps in our knowledge, and about how research might help communities of all kinds.

Table of Contents

List of Figures

List of Tables

Executive Summary

Scope

This document establishes a Federal roadmap in the area of information integrity research and development. Its purpose is to acknowledge gaps in today's scientific knowledge and equip research communities with a shared understanding of Federal research priorities for the future. These priorities enable researchers within and outside of the government to align research efforts, identify opportunities for productive coordination, and work with communities of all kinds to explore the area of information integrity in a manner that protects fundamental freedoms and human rights.

Importantly, this document is not a policy directive. It makes no determinations or recommendations regarding laws or regulations that govern the digital information ecosystems, and it does not take any policy positions. Instead, its purpose is to stimulate research that can be used to strengthen the role of information ecosystems in the open exchange of ideas where healthy debate and free expression thrive.

Background

Accurate and reliable information is central to the well-being of people and society. Manipulated information can have destabilizing and harmful consequences for national security, democratic processes, economic welfare and development, the environment and natural ecosystems, local and national crisis response efforts, human rights and protections, violence and extremism at home and abroad, healthcare systems, and people's individual health and welfare. Rapid technological advances, genuine uncertainties, determined foreign state adversaries or competitors, and internet-scale human-machine interactions create complex dynamics that compound these risks.

Digital interconnectivity now enables billions of people to exchange information at unprecedented speed and scale. Now, many individuals can access hundreds, thousands, or even millions of people, which changes the way information flows and how traditional sources of high-integrity information operate. On the one hand, large-scale connectivity enables new mechanisms to build support systems, exchange ideas, and achieve collective goals; on the other hand, when individuals, companies, and institutions lack the means to discern trustworthy from manipulated information to make sound decisions, opportunities emerge for fraud, harassment, exploitation, safety hazards, crime, health risks, and other types of harm. Although manipulation of information is not new, its spread and effects have been significantly exacerbated by the modern digital landscape and by dynamics that enable manipulated information to flourish. New and powerful AI techniques can be used to generate deceptive content that can be difficult to detect, platforms may amplify corrupted content, and people may inadvertently spread manipulated narratives.

Further, harms associated with manipulated information affect people of all ages, nationalities, and creeds; people of all cultural, ethnic, and racial backgrounds; people of all genders; and people across the political spectrum. The exact nature of harm varies. Some groups are at higher risk, including older Americans, who experienced $184 million of losses from fraud in 2018 alone,[1] and veterans, including military retirees, who reported $66 million of fraud to the Federal Trade Commission in 2020.[2] Others may experience more severe effects from online harassment or abuse, which disproportionally affects women; adolescent girls; LGBTQI+[3] individuals; and those who are additionally targeted because of their race, ethnicity, religion, or other factors. Understanding the risk factors, prevalence, and nature of harm is needed to explore and validate strategies to mitigate risks and address the potential for harm.

It is important to acknowledge the complex work that people in social media companies, news organizations, and parts of the government do to provide high-integrity information to the public. It is also important to acknowledge that, in some instances, these efforts fall short and at times exacerbate misconceptions, raise concerns about free speech, or contribute to unintended consequences. Scoping the array of such possible consequences, while also acknowledging that information ecosystems are dynamic, complex, and at points unpredictable, is critical to mitigating harm.

Recognizing these complexities, and the wide array of risks, the Biden-Harris Administration is unequivocal that preserving democratic principles and protecting public safety necessitates researching information manipulation that has the potential to cause harm, while steadfastly safeguarding American civil rights and liberties, including First Amendment rights. Furthermore, prior administrations and bipartisan coalitions in Congress have concurred that these issues are crucial, taking important actions to investigate and address risks associated with information manipulation. Under the leadership of the White House Office of Science and Technology Policy, in conjunction with the National Science and Technology Council, and in coordination with the National Security Council and Domestic Policy Council, an interagency working group (IWG) was formed. The IWG, together with a diverse array of Federal departments and agencies, collected information from the public on these issues through a Request for Information (RFI) and received briefings from dozens of members of government, private sector, research groups, and community organizations. Its primary purpose is to better understand the risks associated with manipulated information, the complexities of communicating high-integrity information in the context of genuine uncertainty, gaps in knowledge and technical capabilities, and strategies to safeguard civil rights and liberties.

As the result of a year-long study, this document lays out a Roadmap for Researchers on Priorities Related to Information Integrity Research and Development to understand how the lack of high-integrity information contributes to harmful outcomes and what research can help address these challenges. This Roadmap establishes objectives for federally funded information integrity research (both extramural and intramural), provides a structure for coordinating research and development (R&D) in information integrity, and encourages multidisciplinary research that recognizes responsibilities in both the public and private sectors. The Roadmap reflects the perspective of government employees across the political spectrum, with recognition that some research questions will benefit from broader community engagement throughout the research process to align research questions with public interests and remain responsive to the public's perspectives on desirable levels of information integrity, methods to enable free expression, and respect for human rights.

This Roadmap makes no determinations or recommendations regarding laws or regulations that govern the digital information ecosystems, and it does not advocate for the use of any particular approach or action to address information integrity challenges. Instead, its purpose is to advance the science and expand the number of evidence-based options that government employees and policymakers, entrepreneurs and companies, individuals and community organizations, educators, nonprofits, and foreign partners have to strengthen the role of information ecosystems in the open exchange of ideas where healthy debate and free expression thrive.

Goals

Information integrity R&D should focus on understanding how people discern between high-integrity (i.e., authentic, accurate, trustworthy, transparent about its vetting) and low-integrity information and information ecosystems, contend with uncertain or ambiguous information, mitigate and recover from the effects of information manipulation, and enjoy the benefits of open information ecosystems (see Section 1.6 and the Glossary for explanations of the terms used in this document). At the same time, research prioritized by this Roadmap must seek to provide understanding for communities to strengthen the integrity of information exchange and American society to preserve fundamental freedoms across information ecosystems. Likewise, it should provide tools for people and systems to operate effectively, even in the presence of manipulated information. To achieve the vision of high-integrity information ecosystems, research should focus on the following goals:

- **Integrity Assessment**: Enhance approaches people use or would like to use to discern between high-integrity and low-integrity information, narratives, and characteristics of information ecosystems.
- **Harm Mitigation**: Identify strategies that could be used to prevent or minimize avoidable harms caused by information manipulation and assess their effectiveness across populations, cultures, communities, and types of harm.
- **Resilience**: Identify skills and strategies that make it easier for people to operate in the context of potentially questionable information, and enhance communication strategies and tools that enable communities to progress toward and maintain appropriate levels of information integrity within open information ecosystems, even in the presence of uncertainty or active manipulation.
- **High-Quality Evidence**: Collect rigorous empirical evidence to evaluate strategies and technologies intended to address information integrity challenges; clearly convey high-quality evidence to decision-makers to inform the development of relevant public policy; organization-level communication processes; and decisions about which technologies to adopt, enhance, or retire.

Research Priorities

To realize these goals, this Roadmap establishes research priorities for fostering and advancing information integrity science and its application to real-world contexts. The research priorities address a broad range of concerns and challenges in information integrity, describing key objectives for fundamental, use-inspired, translational, and outcomes research, as well as strategies to inform policy and practice. It is anticipated that individual Federal agencies will focus on subsets of the research priorities, corresponding to their expertise. Accordingly, contributing agencies will develop and communicate their prioritization in accordance with their focus and mission.

Collectively, the research priorities seek to answer these questions:

1. How do human, social, technical, and systemic elements of information ecosystems affect the integrity of information creation, exchange, and consumption, and how can modeling, measurement, and analysis enhance understanding of the underlying mechanisms?
2. Which safeguards, skills, and strategies assist people who encounter manipulated, misleading, or uncertain information, what are the risks and benefits, and how does effectiveness vary across context, culture, information source, and population?
3. How might technologies make it easier for communities to foster and maintain high-integrity information ecosystems?

4. What approaches and tools help people identify, mitigate, respond to, or proactively debunk manipulated-information campaigns while protecting First Amendment and human rights?
5. How can data access, research infrastructure, and partnerships accelerate and improve the rigor of information integrity research?
6. How can the knowledge and insights produced by information integrity research inform policy and practice in a manner that helps society as a whole?

The following six research priorities seek to address these questions.

Research Priority 1: Analyze, Model, and Measure Information Ecosystems. Advancing the science of information integrity requires models and measurements with which to instantiate, evaluate, iterate, and improve theories of information ecosystems. Computer science, cognitive sciences, economics, epidemiology, linguistics, political science, psychology, social and behavioral science, and other relevant disciplines have developed methods, models, and theories that offer explanations of different aspects of information ecosystems. A challenge for advancing the science of information integrity is to connect relevant methods, models, and theories to capture interdependencies in the system and explain new conditions as they arise, including mechanisms related to information supply chains, incentive structures, human elements, and technical components. Interdisciplinary theory is needed to advance scientific understanding and formulate research questions that will lead to effective and efficient ways to mitigate harms, bolster information integrity, and strengthen resilience against information manipulation. Significant advances in measurement are required to understand information ecosystems, measure dynamics across media platforms, explore how information ecosystems connect distinct populations, and evaluate strategies to mitigate manipulated information.

Research Priority 2: Investigate Safeguards that Assist People. Manipulated information is easy to produce, is free to distribute on a wide scale, and cannot be easily stopped at its source. Given these realities, research on safeguards that assist people, communities, and organizations—including those who encounter manipulated information and those responding to manipulated-information campaigns—is vital. Research is needed to identify safeguards, skills, and strategies that reduce people's vulnerability to manipulated information and empower people to become astute consumers of information. Importantly, safeguarding approaches must also respect the autonomy of participants in the information ecosystems, protect free expression, and support people in making informed and independent assessments of information. Research is needed on opportunities to familiarize individuals and communities with digital, media, and information literacy outside of schools and other formal learning environments and on how to adapt these support mechanisms to individual and community needs.

Research Priority 3: Envision Technical Approaches to Enhance High-Integrity Information Exchange. Design decisions within information ecosystems create both intended and unintended effects that influence the creation, prevalence, and persistence of manipulated information. For example, evidence suggests that hidden design features, such as the number of times content can be reshared or the effect of emoji reactions on content distribution, may affect the prevalence and reach of manipulated information in ways that are invisible to users.[4,5,6] Similarly, platform design elements contribute to perceived polarization.[7] Such empirical findings raise questions about how technologies might better enable the types of information exchange needed for a community to share perspectives and converge on high-integrity conclusions. Key questions include how technology might assist communities in establishing and maintaining high-integrity communication within open information ecosystems, how provenance-tracking technologies might assist in the design of trustworthy information channels, and how user customization might enable people to influence information ecosystems in ways that appeal to users with a genuine interest in dialog in pursuit of high-integrity understanding. Additional research on the causal relationships between ecosystem design and information integrity properties is also needed, including fundamental research on outcome metrics, algorithm design, and the mechanistic relationships between design choices, societal benefits, and harms related to manipulated information.

Research Priority 4: Understand Effective Strategies to Address Manipulated-Information Campaigns. Campaigns to spread manipulated information deserve special attention because of their dynamism and their continuous development of new tactics to circumvent detection and measurement. For example, deep fake technologies make it significantly easier to create false identities, and advances are needed to keep people from falsifying government IDs, foreign actors from impersonating Americans online, and bad actors from defrauding the government or the public. The expansion of global digital connectivity, immersive information technology, and widely accessible digital marketing techniques is enabling a rapid growth of a manipulated-information industry. Campaigns initiated by foreign states, criminal organizations, and other malicious actors, pose complex technical challenges, because of rapidly evolving tactics and falsification techniques. In parallel with the expansion of malicious actors in the information domain, there is a growing need for mechanisms to address information campaigns that result in harm, such as negative health outcomes, financial theft, radicalization, undermining of international alliances, and degradation of democratic processes. Human-machine teaming approaches are needed to enable early warning, proactive outreach, and mitigation of widespread harm, both to improve communication with communities at risk for being targeted and to rapidly adapt to adversary tactics. Effective approaches must address the asymmetry between the low cost of spreading manipulated information and the high cost of anticipating, mitigating, and remediating effects of such manipulated information and campaigns. Research is needed to understand the incentives for information manipulation and how to impose costs on manipulated-information campaigns to deter them from being initiated. Research will also require a multimodal perspective as campaigns move across numerous online platforms, television, print and electronic media, and offline channels. Finally, interdisciplinary approaches remain essential, including artificial intelligence, national security, economics, education science, behavioral and social sciences, and expertise domains associated with populations targeted by manipulated information.

Research Priority 5: Foster Data Access and Partnerships. Significant advances in information integrity research will require comprehensive, longitudinal data on the production, spread (e.g., digital routing), and consumption of low- and high-integrity information across multiple modes, platforms, and channels. Akin to impacts in domains such as public health or finance, research would benefit from consortia that allow many interested parties—platform companies, researchers, regulators, and end users—to aggregate data, tools, methods, and insights in a coordinated, responsible, privacy protective, and transparent way. Such research consortia would need to grapple with issues pertinent to the information integrity ecosystems, such as adequately protecting proprietary information, personally identifiable information, and other types of sensitive data, as well as the risk of deceptive narratives designed to undermine those engaged in research. In addition, such research consortia must account for First Amendment and human rights considerations, and recognize that private companies retain ultimate discretion, subject to appropriate legal constraints, on how they deal with information integrity issues.

Research Priority 6: Connect Research to Policy and Practice. Ultimately, partnerships should expand beyond data transparency and measurement toward improved outcomes, such as greater public awareness of how to mitigate information manipulation. Multiple research approaches are needed with fundamental researchers providing sound measurement techniques and generalizable knowledge, use-inspired researchers de-risking novel prototypes, translational researchers increasing the utility of those prototypes for policymakers and practitioners, and outcomes researchers measuring longitudinal effects in real-world contexts.

This Roadmap closes with an action plan that identifies potential roles in information integrity R&D for public and private organizations and describes concrete actions for the information integrity R&D community and its partners to consider.

1. About the Roadmap for Researchers on Priorities Related to Information Integrity Research and Development

1.1 Purpose of the Roadmap

The purpose of this Roadmap for Researchers on Priorities Related to Information Integrity Research and Development is to catalyze research and development (R&D) to help the United States and its global partners realize the benefits of open information ecosystems while minimizing potential negative societal impact (see Section 1.6 for the definitions of "information integrity" and related terms). This Roadmap establishes objectives for federally funded information integrity research (both extramural and intramural), provides a structure for coordinating R&D in information integrity, and encourages multidisciplinary research that recognizes responsibilities in the public sector, private sector, and civil society. The overarching goal of this Roadmap is to produce knowledge and technologies that will enable government employees, policymakers, researchers, entrepreneurs, businesses, individuals, communities, nonprofit organizations, libraries, museums, other educational institutions, and foreign partners to benefit from scientific insights, to create transformative technological advancements, and to provide meaningful safeguards around information integrity. The research prioritized by this Roadmap must simultaneously provide the means to strengthen the integrity of information and information exchanges and preserve open dialogue within information ecosystems. This Roadmap guides the R&D efforts in this area; its scope does not include potential responses such as regulation.

While information integrity research occurs across the Federal Government, today's efforts are disjointed, and large gaps exist between the research, policy, and practice. Therefore, this Roadmap outlines priority areas for fundamental, use-inspired, translational, and outcomes research. It provides a framework to coordinate research-related efforts, improve efficacy, and increase return on investment. It seeks to establish feedback loops to enable research insights to translate into effective action and for lessons learned in the field to inform future research investments. Such feedback loops require partnerships between Federal research and service agencies,[8] across human- and technology-centered academic disciplines, and between independent researchers and a variety of groups, including local communities, private companies, and the Federal Government. This Roadmap discusses actions, investments, and partnerships to foster innovation and strengthen information integrity safeguards within the broader exchange of ideas in digital space. It also seeks to communicate priority areas to the research community and inspire parallel efforts in the private sector.

1.2 How This Roadmap Was Developed

As a result of the broad, far-reaching, and significant implications of information integrity challenges, the White House Office of Science and Technology Policy (OSTP) and the Federal Subcommittee on Networking and Information Technology Research and Development (NITRD SC) established the Information Integrity Research and Development Interagency Working Group (IIRD IWG). The IWG was announced by the White House in December 2021 in the FACT SHEET: The Biden-Harris Administration is Taking Action to Restore and Strengthen American Democracy.[9] The IIRD IWG provides a forum for interagency planning and coordination for R&D investments in information integrity, identifying research gaps, defining future research directions, and encouraging public-private partnerships in addressing information integrity challenges.

Throughout the drafting of this Roadmap, the IIRD IWG has consulted with academic researchers, commercial entities, government agencies, international partners, those adversely affected by corrupted information, and other interested individuals seeking to address information integrity challenges. These participants collectively shared expertise across the fields of computing, national security, diplomacy, public health, law enforcement, human trafficking, regulation, consumer protection, behavioral and social sciences, economics, journalism, law, and policy. Through a series of meetings, invited talks, and a formal Request for Information,[10] IIRD IWG members have collected insights from representatives of service agencies, research agencies, nonprofit organizations, corporations, libraries, and the academic research community. The following people and organizations were among those who provided valuable perspectives throughout the process:

Federal Agencies and Councils
Air Force Office of Scientific Research
Census Bureau
Centers for Disease Control and Prevention
Central Intelligence Agency
Domestic Policy Council
Defense Advanced Research Projects Agency
Department of Homeland Security
Department of Justice
Department of State
Federal Bureau of Investigation
Food and Drug Administration
Gender Policy Council
Institute of Museum and Library Services
Intelligence Advanced Research Projects Activity
National Institute of Standards and Technology
National Institutes of Health
National Science Foundation
National Security Council
National Security Agency
NITRD National Coordination Office
Office of Naval Research
Office of Science and Technology Policy
Office of the Director of National Intelligence
Office of the Surgeon General
Office of Undersecretary of Defense (Research & Engineering)
U.S. Agency for International Development

Researchers from Academia
Boston University
Brown University
Cornell University
Dartmouth College
Indiana University
New York University
Northeastern University
Princeton University
Rutgers University
Stanford University
University of Miami
University of Texas at Austin
University of Washington
Vanderbilt University

Nonprofits, Corporations, Individuals
#Shepersisted
American Association of Retired Persons
Anchor Change
Anti-Defamation League
Aspen Digital
Black Brilliance Research
Cardenas Strategies
Carnegie Endowment for International Peace
Centre for International Governance Innovation
Coalition for Content Provenance & Authenticity
Computing Community Consortium
Cliff B.
GFI Research
Integrity Institute
IREX
MITRE
Mozilla
National Human Trafficking Hotline
NewsGuard
Ripeta
Sandia National Laboratories
Sift
Special Competitive Studies Project
WITNESS
Patrick W.

1.3 Intended Audiences

This Roadmap outlines interdisciplinary research priorities relevant to government employees and policymakers, researchers and educators, entrepreneurs and businesses, individuals and communities, nonprofits, and foreign partners. It discusses coordinated actions and research priorities for Federal agencies that fund research as well as those agencies that create incentives to encourage independent research in the private sector. Likewise, it identifies use cases of interest to government practitioners to better align research investments with problems encountered in the field. It provides private sector technology developers and policymakers with information about synergies, opportunities for research collaborations, and areas for parallel efforts. Finally, it provides local communities with potential resources to obtain information related to information integrity and to contribute to research efforts.

1.4 Scope

This Roadmap seeks to generate knowledge and technologies to assist government employees and policymakers, researchers and educators, entrepreneurs and businesses, individuals and communities, nonprofits, and foreign partners with studying and addressing information integrity challenges. This Roadmap does not advocate for the use of any particular technical approach; rather, it aims to expand the number of evidence-based options available to the aforementioned individuals and groups. By exploring what techniques work, why they work, and when they fail, this Roadmap aims to develop knowledge and enabling technologies that could inform policy development and implementation.

This Roadmap, while it considers the legal context, makes no determinations or recommendations regarding laws or regulations that govern digital information ecosystems. Similarly, this Roadmap does not make recommendations to service departments and agencies as to what actions to take to address information integrity challenges. Such decisions are the responsibility of the departments and agencies themselves, subject to the authorities granted to them by Congress.

1.5 Relationship to Bipartisan Executive and Congressional Actions

Addressing the challenges associated with information integrity has yielded bipartisan efforts over multiple administrations, including bipartisan legislation, bipartisan interest at Congressional hearings, and extensions of executive actions (supporting documentation can be found in Appendix A). The following bipartisan efforts focus on mitigating threats from foreign adversaries and harms to the American public, and these policy goals help to frame the harm-mitigation emphasis within the Roadmap.

- **Threats from foreign propaganda and disinformation.** Countering disinformation campaigns by America's adversaries has led to notable bipartisan efforts to counter foreign disinformation that aims to undermine U.S. national security interests, confront foreign malign influence, and address covert distribution of disinformation that aims to undermine elections.[11,12,13,14]
- **Protection against scams**. Consumer harms perpetrated through disinformation spread online has been the source of a continued congressional hearings, including investigations into impersonation of veterans' groups that resulted in a reported $66 million in losses among military retirees and veterans,[15] as well as ongoing scams related to COVID-19, such as identity theft, fake personal protective equipment, and fraudulent vaccination claims. [16,17]
- **Harms to children and youth.** A multitude of congressional hearings have focused on the impacts to teens from social media platform recommendation algorithms such as social media's impact in promoting eating disorders among its younger users[18] and the misuse of social media

applications to harm kids, promote destructive acts and deadly challenges, bullying, manipulative influencer marketing, and grooming.[19]

- **Elder fraud.** In 2022, Congress enacted several bipartisan efforts to address elder fraud, prevent scams that target seniors, monitor mail, television, internet, and telemarketing for fraud targeting seniors, and examine mechanisms to assist aging Americans.[20,21,22,23]

In addition to its domestic focus, the Biden-Harris Administration's prioritization of efforts to counter information manipulation has been reinforced by global partners; for example, in the Declaration for the Future of the Internet, launched by the United States with 60 partner countries from around the globe.[24] The Biden-Harris Administration has underscored in its approach on these issues the following priorities:[25, 26]

- **Strengthen democracy at home and abroad:** Defend against foreign interference that is aimed at undermining confidence in U.S. elections, mitigate threats that misinformation and disinformation pose to public safety and national security, and bolster democracy and human rights globally.
- **Protect the right to freedom of expression:** Reaffirm the commitment that actions taken by the U.S. Government and authorities will be consistent with the constitutional right to freedom of expression and international human rights law, while encouraging diversity of opinion and pluralism without fear of censorship, harassment, abuse, intimidation, or the "chilling effect" that leads to self-censorship among victims.
- **Foster resilience:** Support innovative ways that can improve individual and community resistance to misinformation and disinformation, including by supporting media, scientific, health, and digital literacy programs and critical thinking skills.
- **Mitigate extreme social polarization:** Address disinformation and domestic and foreign malign activities used to sow division and conflict between individuals or groups of individuals in society aimed at undermining respect for civil rights and liberties and for democratic institutions.

1.6 Terminology

This section explains several of the key terms used in this Roadmap. The Glossary provides information about and definitions for additional terms used throughout this document.

1.6.1 The Information Integrity Spectrum

In this Roadmap, ***information*** refers to any representation of knowledge or expression of ideas, including facts, opinions, and conjectures, which may reference data, statements, imagery, and sound in any medium or form, including textual, numerical, graphic, cartographic, narrative, audiovisual, or aural.

This Roadmap uses the term ***information integrity*** to describe the spectrum of information and associated patterns of creation, exchange, and consumption in society, where high-integrity information is trustworthy; distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting. This information can be linked to the original source(s) with appropriate evidence. High-integrity information is also accurate and reliable, can be verified and authenticated, has a clear chain of custody, and creates reasonable expectations about when its validity may expire. In cases where indicators of consensus support are available, those indicators are authentic and accurately represent the size and nature of support, with reasonable error bounds. An information ecosystem with high levels of information integrity is one in which people and organizations effectively communicate the best information available, in which threats to information integrity are proactively addressed, and in which individuals and populations are empowered by critical thinking skills to move toward high-integrity conclusions about the quality and intent of information they encounter.

Threats to ecosystem information integrity are complex phenomena. Terminology varies because of the diversity in threat source and intent (see the Glossary and the following bullets for explanations of the commonly used terms, including *misinformation*, *disinformation*, and *malinformation*). This Roadmap is interested in two dimensions of low-integrity information: the potential to cause harm and the intent to deceive. This Roadmap uses the term ***corrupted information***[27] as an umbrella term that includes erroneous information that may accidentally cause harm if presented in the wrong context (also referred to as *harmful misinformation*) and information characterized by intentionally deceptive elements that may lead to harm (also referred to as *manipulated information*).

Corrupted information encompasses information that is inaccurate, misleading, deceptive, or inauthentically amplified within the context presented to the information consumer and that may cause harm to individuals, communities, institutions, or society at large, regardless of the author's or the distributor's intent. ***Manipulated information*** describes the subset of corrupted information that is produced or altered with an intent to deceive, mislead, or attack for some type of gain. Manipulation may include changes to the information itself, to the context, as well as distorting amplification[28] to deceive audiences about the volume, balance, or public appeal of the information.

Historically, researchers and practitioners have used an array of terms that communicate intent and distinguish the presence of false information from the use of accurate information to imply a faulty conclusion.[29] Common terms include the following:

- **Misinformation:** False, inaccurate, or incorrect information that departs from the facts or preponderance of evidence; it may be innocuous. Misinformation is also defined as false information not created or shared with the intention of causing harm.
- **Disinformation:** Information that contains false content and is created or spread intentionally with the purpose of altering a specific target audience's attitudes or behavior.
- **Malinformation:** Information that is accurate but is strategically spread or slanted to cause harm.

This Roadmap centers on two terms: *information integrity* and a new term, *corrupted information*. The principal reason for using *corrupted information* is that there is currently no umbrella term that focuses specifically on the potential to cause harm, regardless of intent. In addition, there is no umbrella term that explicitly encompasses both aspects of the content itself as well as deceptions around its social context (such as falsified amplification techniques or deceptions regarding who or how many people adhere to an idea).

1.6.2 Open Information Ecosystems

The term ***information ecosystem*** refers to a dynamic and evolving set of entities and technologies that collect, pass, disseminate, act on, or otherwise engage with information content. An ecosystem is composed of information producers and consumers, groups and social networks, cultural influences, technology, market forces, system dynamics, and social institutions and contexts, in which entities may fill roles that overlap and change over time. Information ecosystems include human, technical, and sociotechnical elements, where the term ***sociotechnical*** denotes complex interactions between humans and technology that cannot be well characterized by studying either the human or technical components in isolation. Sociotechnical complexity, for instance, could stem from the interaction between artificial intelligence (AI) and social systems in which thousands or millions of human actions interact with automated systems in ways that may be opaque to both system designers and users. In this Roadmap, the term ***sociotechnical*** emphasizes areas where research would benefit from interdisciplinary collaborations between technology researchers, social scientists, and behavioral scientists to investigate complex phenomena. Appendix B contains a more detailed treatment of these ecosystems.

An ***open information ecosystem*** enables the free exchange of ideas and use of technologies. It supports a free exchange of ideas, enables ideas to flow from multiple sources, empowers people to express conflicting perspectives in a constructive manner, and leverages a free market of technologies to distribute information to audiences.

2. Strategic Framing

2.1 Risks and Benefits of Open Information Ecosystems

Digital interconnectivity now enables billions of people to exchange information at unprecedented speed and scale. With today's technologies, many individuals can access hundreds, thousands, or even millions of people, changing the way information flows, how traditional sources of information integrity operate, and levels of trust in institutions that historically provide high-integrity information. On the one hand, large-scale connectivity enables new mechanisms to build support systems, exchange ideas, and achieve collective goals; on the other hand, when people lack the information or trust in credible sources that they need to make sound decisions, opportunities emerge for fraud, harassment, exploitation, safety hazards, crime, and other types of harms.

Although manipulation of information is not new, its spread and effects have been significantly exacerbated by the modern digital landscape. AI techniques can generate deceptive content that is difficult to detect, platforms amplify corrupted content, and people inadvertently spread manipulated narratives. Moreover, the greatest impact may stem from online technologies' capabilities to rapidly disseminate manipulated information on an exponential scale. The ability of media platforms to reach over a third of the world's population, and the lack of consistently applied standards and ethics in the production, propagation, social listening, and response to corrupted information, can block efforts to inform citizens responsibly and impede efforts to address harmfully corrupted information.

Innovation also creates the possibility of incorporating new safeguards into open information ecosystems, allowing people to reimagine the techniques and technologies used to address corrupted information. Knowledge, creativity, participatory research, and sustained investments are needed at the touchpoint between people, technology, and the free exchange of ideas to maximize the benefits of open information ecosystems and to mitigate harms.

2.2 Harms from Corrupted Information

Both corrupted and manipulated information have the potential to inflict serious harm to people or society, which in turn discourages free expression and affects the security of the United States. Examples include threats of violence, financial fraud, debasement of individuals or groups, damage to public health and safety, interference with public services during disasters, and harms to national security. Mitigating such harms and addressing factors within information ecosystems that increase risks to people and populations are among the key objectives of this Roadmap. Appendix C contains a full description of the harms considered in this report with extensive discussion and sourcing. The following text provides a summary of these harms, which are listed in alphabetical order.

2.2.1 Harms to Consumers and Companies

Financial harm to consumers. Financial harm from deceptive ads and other manipulated information on social media platforms is a growing problem, creating substantial risk to consumers.

Harms to corporations. State and non-state actors are using corrupted information to wage economic warfare or manipulate consumer confidence in markets.

Service disruptions. False claims can lead to destruction of corporate assets, damage to critical infrastructure, and service disruptions.

2.2.2 Harms to Individuals and Families

Elder fraud. The number of older Americans falling victim to internet crimes such as romance scams, investment fraud, healthcare miracle cures, government impersonation, and tech support fraud has risen dramatically with the rise of broadband connectivity, reaching $184 million of losses in 2018 alone.[30]

Harms to youth. Children and youth constitute a population especially vulnerable to corrupted information, and emerging research raises acute concerns about the damaging effects of false and misleading information on children.

Health. Corrupted health information can contribute to physical and mental distress, injury, disability, or death, especially when it causes individuals to refuse life-saving treatments or take actions detrimental to their own health.

Personal agency. By design, information manipulation seeks to alter a person's beliefs and actions. Harm is often created when those beliefs or actions are not what the person would have chosen in the absence of the manipulation.

Safety and reputation. Broad safety concerns, such as harassment, child exploitation, cyberbullying, and violent extremism, in addition to reputation concerns, defamation, invasion of privacy, politically motivated harassment, and abuse that disproportionally affects women, are among the key issues motivating research in information manipulation.

2.2.3 Harms to National Security

National preparedness. Corrupted information may have significantly destructive impacts on the state of U.S. preparedness to sustain national critical infrastructures and respond to rapidly evolving events such as disasters and emergencies.

Rule of law. Corrupted information compounds such already onerous rule-of-law challenges as human trafficking, illegal immigration, and terrorism by straining limited resources and inducing the participation of individuals who might not otherwise engage in criminal activity.

Strategic miscalculation. Information advantage is fundamental to military preparedness and sound decision making. National security policies stemming from corrupted information, whether as a function of purposely falsified intelligence or of the outraged demands for action from a public that has been deliberately misled by a nation state adversary, can be damaging to a nation's strategic interests.

Widespread civil unrest. Corrupted information has been employed globally as a means of attempting to destabilize and fracture societies along political, religious, and racial fissures.[31]

2.2.4 Harms to Society and Democratic Processes

Correlated escalation from online content to offline violence. Online manipulated information that may motivate offline violence has been identified as a factor in tragedies such as genocide, attacks on infrastructure, or targeted attacks on specific populations. [32,33,34]

Distrust in elections. Because democratic satisfaction and participation are closely associated with the perceived legitimacy of election outcomes, actual and rumored information manipulation over elections poses a significant threat to democracy.[35]

Distrust of the media and threats to free press. The ability to both falsify information and present it in formats that look or act like news can make it harder for people to distinguish between sources and articles that follow journalistic standards and those that do not.

Oppression of women and other marginalized groups. Manipulated information on social media often involves targeted harassment campaigns that seek to silence and marginalize specific groups in society, and make it appear that actors spreading manipulated information are conveying an outsized consensus.

Polarization. Information manipulation is used to reinforce insular and mutually suspicious online communities in which false claims are repeated, magnified, and considered as facts, making discourse between different groups of people more difficult. Once established, polarization provides fertile ground for social manipulation.[36]

2.3 Desired Outcomes

Trustworthy information plays a critical role in a well-functioning democracy and society, and communication that leads to reliable conclusions is vital to government employees and policymakers, researchers and educators, entrepreneurs and businesses, individuals and community organizations, nonprofits, and foreign partners. Therefore, the research prioritized by this Roadmap must simultaneously seek to strengthen the integrity of information exchange and to preserve open dialogue within information ecosystems. The following text discusses the desired outcomes that should provide the vision for future research to strengthen information integrity.

2.3.1 Toward Constructive Public Debate

Public discourse is at the heart of a free and open society. It has changed with technological advances that have contributed to a fractured perception of reality as people create, manipulate, exploit, share, and consume information in novel ways. There is a need for sociotechnical mechanisms that can enable and support constructive interpersonal dialogue toward a shared understanding of complex topics in a manner that bridges divisions.

2.3.2 Toward a Well-Functioning Marketplace of Ideas

The "marketplace of ideas" is an enduring sociopolitical concept that free and unfettered communication and exchange of ideas is emancipatory and essential for well-functioning democracies.[37] This aspirational concept borrows from free market theories and is an idealized notion that the competition of ideas will result in the accurate and high-integrity ideas gaining acceptance and the low-integrity alternatives being rejected. The claim that free exchange will converge on high-integrity outcomes continues to be a topic of debate. Nonetheless, the salient ideas, that all points of view should be afforded an opportunity to be heard and that free speech is essential in the development of informed choice, should be protected. Correspondingly, new mechanisms are needed to support high-integrity outcomes within the marketplace of ideas, while protecting both First Amendment and human rights.

2.3.3 Preservation of Democratic Principles

Civic participation is a goal of democracies worldwide. It can be undermined by threats to a free press, foreign interference, and popular distrust of election processes. Healthy information ecosystems can reduce these threats, especially if they are designed with mechanisms that support both civic participation and open, transparent communication between candidates, representatives, and their constituents.

2.3.4 Toward Trust in Credible Sources of Information

When faced with uncertainty, people often look to trusted individuals, organizations, or institutions for perspective and information. As trust in government, the free press, academia, and civic institutions continues to erode, it leaves a void that affects both where people go to find information and the types of information sources they are willing to trust. Techniques are needed to help people identify credible sources of information that are worthy of their trust and to reestablish trust with national institutions whose employees are actively working to strengthen connections with local communities and provide the public with high-integrity information.

2.3.5 Toward Trust in Evidence-Based Enterprises

Evidence-based enterprises, including but not limited to science, medicine, journalism, and education, collectively aspire to explore areas of high uncertainty; develop conjectures that may later be proved false; create standards to reduce the prevalence of fraud; communicate findings of the best available research; and may take months, years, or decades to converge on an accurate understanding of complex phenomena. These characteristics contribute to the inadvertent generation of incomplete or contradictory knowledge that can lead to the spread of mistaken information. Better understanding is needed of how to preserve trust around these endeavors, when they are trustworthy, and how to educate and communicate with the public about the beneficial results of multi-generation evidence-based endeavors along with the uncertainties inherent in exploratory processes. Preserving trust in evidence-based enterprises should also recognize and address the impact of scientific fraud (e.g., the use of false scientific information to advance an agenda or personal interests), publication biases, and the challenge of replicating results.

2.3.6 Toward Sound Decision-Making in the Presence of Corrupted Information

Tolerance for corrupted information is a necessary part of living in a free and open society. Nonetheless, individuals, communities, and organizations would benefit from evidence-based strategies and tools to help them make sound decisions, even when the information landscape becomes complex. Regardless of whether people are making decisions that affect their own autonomy or deciding whether to act or exercise restraint in response to corrupted information, effective techniques are needed to help people consider the impact of the information used in their decisions.

2.4 Goals for Information Integrity Research

Information integrity R&D should focus on creating knowledge, technologies, and sociotechnical insights that will enable people to discern between high-integrity and low-integrity information and information ecosystems, contend with uncertain or ambiguous information, mitigate and recover from the effects of information corruption, proactively fill information voids, and enjoy the benefits of open information ecosystems.

To advance knowledge and technical innovation, information integrity research should focus on the following equally important goals:

- **Integrity assessment:** Enhance approaches people use or would like to use to discern between high-integrity and low-integrity information, narratives, and characteristics of information ecosystems.
- **Harm mitigation:** Identify strategies that could be used to prevent or minimize avoidable harms caused by corrupted information and assess their effectiveness across populations, cultures, communities, and types of harm, with additional attention to historically underserved groups.

- **Resilience:** Identify skills and strategies that make it easier for people to operate in the context of potentially questionable information, and enhance communication strategies and tools that enable communities to progress toward and maintain appropriate levels of information integrity within open information ecosystems, even in the presence of uncertainty or active manipulation.
- **High-quality evidence:** Collect rigorous empirical evidence to evaluate strategies and technologies intended to address information integrity challenges; clearly convey high-quality evidence to decision-makers to inform the development of relevant public policy; organization-level communication processes; and decisions about which technologies to adopt, enhance, or retire.

2.4.1 Integrity Assessment

Enhance approaches people use or would like to use to discern between high-integrity and low-integrity information, narratives, and characteristics of information ecosystems.

Assessing the integrity of information is a complex and multifaceted problem with many technical challenges and subjective aspects. Digital information can be altered, misrepresented, faked, or embedded with hidden content to deceive or mislead. Many gaps exist in the current capabilities to rapidly detect information corruption and identify manipulators across different forms of media (e.g., text, audio, image, video, and virtual/augmented reality data). Gaps also exist in understanding how social, cognitive, emotional, behavioral, and cultural factors affect people's assessment of information integrity. In addition, deeper understanding is needed of how content, its dissemination channels, other actors, sources, and perceived motivations affect people's assessment of both low- and high-integrity information.

One of the key goals for research is to develop approaches to discern between high-integrity and low-integrity information, narratives, data, and characteristics of information ecosystems. Assessment of information ecosystems would capture the effects of technical infrastructure design and social dynamics on platform-wide integrity levels, such as the prevalence, accessibility, and visibility of high-integrity versus low-integrity information inside the ecosystem. Such approaches should be able to establish indicators that place items on an integrity spectrum and distinguish between objective and subjective dimensions of integrity, as in the following examples:

- Whether a piece of information or narrative is authentic with clear provenance.
- Whether the content of the information or narrative reflects objectively verifiable facts and evidence.
- Whether distinctions between facts, opinions, inferences, and uncertainties are clear.
- Whether the information or narrative is designed to evoke an emotional, behavioral, or social response that causes recipients to engage quickly without thinking critically about integrity indicators.
- Whether the source of the information or narrative is trustworthy or credible and whether the dissemination remains free of intent to deceive or harm.
- Whether there is evidence of bots fabricating responses to content to deceive others about the levels of popular support for that content or to illegitimately amplify the content.

2.4.2 Harm Mitigation

Identify strategies that could be used to prevent or minimize avoidable harms caused by corrupted information and assess their effectiveness across populations, cultures, communities, and types of harm.

This goal focuses on corrupted information that causes significant harm to individuals or society, especially when that harm is avoidable. Serious harmful outcomes warrant special attention to understand causal mechanisms, communities at risk, and the scale of the problem as well as evidence-based strategies to diffuse corrupted information campaigns. Participatory research models are especially important in this area because the perception of harm and risk may sometimes be subjective. It therefore may be practical to use public advisory boards to help prioritize the harms to address, and to focus evolving research efforts, as well as to inform the types of actions that may be appropriate in response to evidence produced by research. Forming engaged, participatory approaches such as these will require the establishment of sustained, mutually beneficial partnerships with impacted communities, and may also necessitate the development of research incentives that align with public interests rather than with publication records.

As noted in Section 2.2, the range of potential harms is broad, to individuals, to companies, and to society. Likewise, the risk factors and prevalence of harmful outcomes vary across different populations, cultures, and social groups. Nonetheless, the direct and indirect harms from corrupted information are difficult to quantify. A key point of this Roadmap is that, to understand the effects of corrupted information, the causal mechanisms, and strategies to mitigate the harms, it is important to understand communication both as the transmission of information and as a mechanism for communicating beliefs and group affinities. Therefore, this Roadmap advocates understanding of the social and psychological roots of corrupted information to inform the development of harm-mitigation techniques. This Roadmap also suggests that exploring and examining harms may prove a productive approach for scoping the problem and identifying use cases that can lead to more generalizable findings across methods and research disciplines. Approaching the challenges of manipulated information from a use case perspective, and rallying cross-disciplinary attention to harm mitigation, may help to break down traditional disciplinary silos.

2.4.3 Resilience

Identify skills and strategies that empower people to operate in the context of potentially questionable information, and develop communication strategies and tools for open information ecosystems that enable communities to progress toward and maintain appropriate levels of information integrity, even in the presence of uncertainty or active manipulation.

Resilience is generally defined as the ability to operate effectively in the face of adversity, to withstand and rapidly recover from disruptions, and to adapt to changing conditions.[38] Resilience provides an organizing framework for research—to holistically consider the technological, social, cultural, and policy components of the information ecosystem; the types of required actions (identify, anticipate, ignore, respond, recover, adapt); and the desired future state. Within the information ecosystem, resilience to information corruption applies to people, systems, and policies.

Increasing societal resilience to information manipulation requires an understanding of how certain communities are targeted; what factors influence people’s susceptibility to manipulated narratives; how manipulated information is propagated; and the impact of information manipulation on people's beliefs, thinking, attitudes, and behaviors. A key element of strengthening resilience against information manipulation is to empower individuals through education on how to recognize, create, consume, and propagate trustworthy information and to identify corrupted information. Effective educational pathways are needed for all age levels, demographics, technological experiences, and life circumstances.

Strengthening resilience also requires building an information infrastructure that is resistant to information corruption. Building a resilient information infrastructure will require developing effective capabilities to understand strategies and tactics used to generate and disseminate corrupted information; designing mechanisms for assessing, presenting, and signaling trustworthy information; establishing robust analysis techniques to discover evolving forms of manipulation; and identifying methods to adapt to evolving manipulation strategies.

2.4.4 High-Quality Evidence

Collect rigorous empirical evidence to evaluate strategies and technologies intended to address information integrity challenges; clearly convey high-quality evidence to decision-makers to inform the development of relevant public policy; organization-level communication processes; and decisions about which technologies to adopt, enhance, or retire.

A priority should be to create generalizable knowledge from which to build evidence-based policy and practice.[39] A vast gap currently exists between rigorous research evidence and the policymaking and practice that should be informed by this evidence, so it is important to understand how research can inform and strengthen the policymaking process and the development of more effective tools for practitioners. This will require, at a minimum, an understanding of specific issues or problems that might reasonably be addressed by policy or practice; attention to adequately translating empirical knowledge into actionable options, trade-offs, and recommendations for policymakers or practitioners; and commitment to conducting formal, empirical evaluation of proposed evidence-based policies or practices from the start, including consideration of their potential impacts, whether intentional or not.

3. Information Integrity Research Priorities

Research toward integrity assessment, harm reduction, improved evidence, and resilience must address gaps in knowledge and capabilities. Progress requires a deeper understanding of integrity-related phenomena as well as explorations into when and why strategies to diffuse corrupted information succeed or fail. However, integrity challenges vary across components of an information ecosystem. Therefore, this Roadmap separates research on fundamental aspects from challenges specific to information recipients, technological elements, and corrupted information campaigns. The research priorities outlined in this section address foundational work (models and measurements of information ecosystems), people-oriented support mechanisms (safeguards that assist people), technology-oriented approaches (technical approaches to enhance high-integrity information exchange), and operationally-oriented tools (effective strategies to address manipulated-information campaigns). Figure 1 illustrates how the four research priority areas interconnect and collectively help achieve high-integrity information ecosystems.

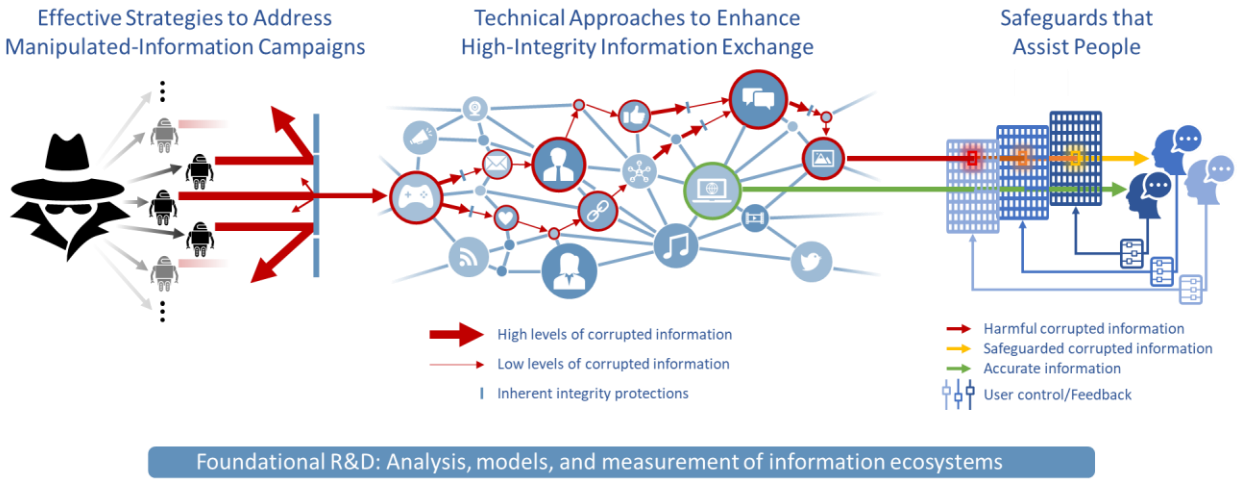
Effective Strategies to Address Manipulated-Information Campaigns
Technical Approaches to Enhance High-Integrity Information Exchange
Safeguards that Assist People
High levels of corrupted information
Low levels of corrupted information
Inherent integrity protections
Harmful corrupted information
Safeguarded corrupted information
Accurate information
User control/Feedback
Foundational R&D: Analysis, models, and measurement of information ecosystems

Figure 1. Depiction of Core Information Integrity Research Priority Areas

Table 1 depicts the relationship between these priorities and the information integrity goals defined in this Roadmap. Fundamental, use-inspired, translational, and outcomes research will all play key roles, with system-level explorations focusing primarily on generalizable knowledge, and the remaining priorities expanding into translational and outcome-centric efforts.

Table 1. Overview of How the Research Priorities Will Address the Information Integrity Goals

	Information Integrity Goals			
Research Priorities	Integrity Assessment	Harm Mitigation	Resilience	High-Quality Evidence
Analyze, Model, and Measure Information Ecosystems	Advance the science of information ecosystem measurement	Understand the human dimensions of corrupted information	Develop system-level models of information ecosystems	Advance the science of information ecosystem measurement

	Information Integrity Goals			
Research Priorities	**Integrity Assessment**	**Harm Mitigation**	**Resilience**	**High-Quality Evidence**
Investigate Safeguards that Assist People	Enhance education on media, digital, and information literacy	Establish safeguards for specific instances of corrupted information	Enhance education on media, digital, and information literacy	Assess how information sources and cultural contexts impact safeguard effectiveness
Envision Technical Approaches to Enhance High-Integrity Information Exchange	Develop approaches for socially responsible innovation	Design trustworthy information channels	Innovate to support constructive engagement around information	Develop approaches for socially responsible innovation
Understand Effective Strategies to Address Manipulated-Information Campaigns	Anticipate harmful information campaigns	Develop deterrence techniques for foreign state manipulations and illegal activity; Develop proactive strategies to deescalate campaigns	Develop proactive and effective strategies to deescalate campaigns	Enable outcomes monitoring

It is important to acknowledge that information manipulation is a rapidly moving target, and current capabilities will have to be adapted as new threats emerge. To tackle these hard problems reliably and at speed and scale, disciplinary and interdisciplinary research is needed that draws from many research areas, including computer science, engineering, education, information sciences, economic sciences, and behavioral and social sciences.

The following subsections organize the research priorities as follows:

- **Priority Areas:** Identify key, high-level challenges and needs that should be addressed by information integrity research
  - **Research Objectives:** Describe specific research objectives within a priority area
    - **Research Actions:** Provide succinct descriptions of research deliverables that are expected to be accomplished to achieve a research objective
  - **Measuring Progress:** Identify ways of assessing the progress toward achieving the objectives within a priority area

3.1 Priority Area: Analyze, Model, and Measure Information Ecosystems

This Roadmap highlights several recommended focal areas for fundamental research pursued to advance the science of information integrity: (1) developing theories and models, (2) understanding the human dimension of corrupted information, and (3) developing measurement techniques to enhance understanding of information ecosystems.

3.1.1 Research Objective: Develop System-Level Models of Information Ecosystems

Computer science, behavioral and social sciences, economics, epidemiology, linguistics, political science, psychology, sociology, library and information sciences, and other relevant disciplines have developed methods, models, and theories that offer explanations of different aspects of information ecosystems. A challenge for advancing the science of information integrity is to connect relevant methods, models, and theories to capture interdependencies within information ecosystems and explain new conditions as they arise. Transdisciplinary theory is needed to advance scientific understanding and formulate research questions that will lead to effective harm mitigation, evidence-based strategies, and resilient technologies and populations.

Macro-, meso-, and micro-scale models are needed to capture complex ecosystem dynamics and the effects of rapidly evolving threats, embedded AI methods, and internet-scale social interactions. Several organizing frameworks have been introduced,[40] but combining descriptive features with explanatory and predictive capabilities remains a challenge. Descriptive frameworks require theory to progress from disjointed and fragmented investigations to scientifically grounded conceptual and computational models. Ultimately, models should enable impact evaluation to explore how actions, events, and support mechanisms might affect information integrity within an ecosystem.

Incentives around corrupted information. Advances in economic, game, attention, and motivation theories are needed to model the effects of incentives structures on content generation and spread. The online advertisement economy is built on payments that compete for user attention, creating markets for attention-grabbing, even if false, content. Because content with corrupted information can often be made attractive relative to other content, the economic incentives around attention and the tools used to direct it can lead to exacerbated risks of corrupted information. Furthermore, starting a new website or video stream is substantially easier than establishing a new newspaper or radio station, which lowers the barrier to entry for those who wish to distribute corrupted information. Understanding these and other incentives; the role of attention; the role of algorithms, platforms, and click-bait tactics in steering attention; and the underlying economics is an important area for research, as are techniques for adjusting reward structures in ways that incentivize the creation and exchange of high-integrity information.

Information supply and dissemination. The rapid dissemination of corrupted information, whether for gain, harm, or entertainment, can discourage or even overwhelm the free exchange of high-integrity information. Theories are needed to understand how human, technology, and sociotechnical factors influence speed, scale, and spread of corrupted information as well as the role played by influencers, corrupted information campaigns, and nation states that impede the flow of high-integrity information. Also, comparative analysis is needed to understand how the dissemination and supply of high-integrity information differs from that of corrupted information around the world. Special attention should be paid to models that consider multiple channels, as corrupted information often originates from multiple sources and moves across many media platforms, social platforms, international borders, cultures, and languages. Although corrupted information sometimes resonates system wide, linguistic differences can obfuscate corrupted information that targets specific subcultures, and much remains unknown about dissemination, spread, and susceptibility within subpopulations. Deeper understanding of supply chain interactions, key players, channels, and cross-cultural differences is needed, including the risks posed by regional differences in platform safeguards and underinvestment in information integrity research in some parts of the world.

Threats, tactics, techniques, and procedures. Successful campaigns may involve tactics to build and cultivate an audience, introduce corrupted information at an appropriate rate, and craft corrupted information into narratives designed to change beliefs in an enduring way. Malicious actors may cultivate audiences using bots or semiautomated tactics to influence platform amplification algorithms. Information manipulated by foreign state actors deserves special attention, especially regarding the tactics, tools, and strategies that state-controlled entities may use to influence public discourse. More broadly, research is needed to better understand the interplay between content creation, automation, human behavior, algorithmic responses, and outcomes to elucidate sociotechnical campaign tactics and techniques.

3.1.1.1 Research Actions: Develop System-Level Models of Information Ecosystems

- Characterize, explain, and model the interaction of entities, threats, processes, and technologies that affect information integrity, including the psychological, behavioral, and cultural aspects.
- Model the tactics, techniques, procedures, effects, and supply chains for manipulated information perpetrated through campaigns or by foreign state actors, or both.
- Characterize regional and cultural variations in campaign tactics, strategies, threat models, narratives, and effects.
- Develop and validate mechanistic pathways or causal models around harms, high-speed dissemination of manipulated information, and systemic factors that improve integrity properties.
- Develop models that enable impact evaluation of incentives and disincentives for information manipulation and their effects on diverse actors, including attackers, advertisers, media outlets, and social media platforms.

3.1.2 Research Objective: Understand the Human Dimensions of Corrupted Information

It is critical to understand what makes people susceptible to believing, propagating, and acting upon corrupted information, as well as why some people engage in creating and disseminating it. This understanding should include social, cognitive, emotional, cultural, behavioral, and economic aspects and the role that belief systems, echo chambers, cognitive biases, stress, and group association play, among many other variables that may change as technology advances. Being able to identify and measure the factors that make such information persuasive and the factors that make people vulnerable to being manipulated will provide a basis for understanding the causal connections between manipulation and harm.

To advance the understanding of the human dimensions, priority should be given to increasing understanding in the following areas: the bases of beliefs and belief change; how people process new information; the nature, causes, and effects of widely distributed harmful narratives; and the social and psychological aspects that affect the spread and persistence of corrupted information at the group level.

Bases of beliefs and belief change. Interdisciplinary research in behavioral, cognitive, social, cultural, communication, and human-computer interaction sciences is aimed at explaining why people believe and spread corrupted information, and at predicting the spread, evolution, and outcomes of corrupted information. For example, research has already indicated that group loyalty is a major predictor of sustained belief in corrupted information; people copy the beliefs and behaviors of others in their group, and those common beliefs, in turn, reinforce their group allegiance.[41,42,43]

How people process scientific evidence. A priority for research should be to understand why a considerable minority of the public globally mistrusts the findings of science, medicine, and reputable sources of information, and what factors make people more open to learning and investigating further. The reasons for mistrust and disbelief in the findings of science include, but are not limited to, fear (such as of disease or government), facts diverging from ideology, and vested interests (including financial).[44] Factors such as lived experience, interpersonal trust, the understandability of the presentation, and the nature of the process through which scientists create and resolve legitimate uncertainties, may also play a role.

Understanding social and community effects. Corrupted information is not only an individual, but also a distinctly social, phenomenon. Research should explore the social, cultural, and behavioral aspects that cause corrupted information to take hold and endure within a community. It should further examine how group dynamics affect individuals' disposition toward high-integrity versus low-integrity information on certain topics, and how group effects and identity impact both harm and harm mitigation. Likewise, relationships between a person's information ecosystem and his or her risk of health or safety harms also warrant attention. Finally, approaches are needed that would enable people to talk about instances of corrupted information without further propagating it or unintentionally expanding the number of people who believe low-integrity narratives.

Understanding polarization. Evidence suggests that social media use intensifies divisiveness and exacerbates underlying societal divisions,[45] creating risks for meaningful dialogue, shared understanding, and the type of civic discourse needed for well-functioning democracies around the world.[46] The role of media platform algorithms and their interactions with social, cognitive, emotional, cultural, behavioral, and economic factors warrant further investigation to explore whether algorithms reinforce polarization by affecting attitudes or by focusing on delivering more content similar to what people have already engaged with positively. This technical dimension of media platforms both reinforces and amplifies human behavioral tendencies to seek out similarity and avoid difference,[47] and may function to press people away from more critical or rational information-processing approaches in favor of emotionally charged decision-making.[48] It is critical to understand the relationships between human cognition, algorithms, information integrity, and polarization.

3.1.2.1 Research Actions: Understand the Human Dimensions of Corrupted Information

- Develop reliable methodologies to measure persuasiveness of corrupted information and what causes it to affect cognition, beliefs, decision making, and behavior.
- Explore the role of memory, emotion, cognitive biases, reasoning, and social group association in the development of beliefs and belief change, especially related to beliefs about science, medicine, and other sources of evidence-backed information.
- Identify and explain the factors and causes that create false narratives and their spread and consequences.
- Analyze the impacts across the information ecosystems on the cognitive, social, emotional, and health outcomes of individuals, such as exploring which factors constitute social determinants of health.
- Advance the understanding of the dynamics of polarization, including the relationship between polarization and cognitive processes such as identity development and confirmation bias, and the contributions of media platform mechanisms to polarization.

3.1.3 Research Objective: Advance the Science of Information Ecosystem Measurement

Significant advances in measurement are required to understand the information ecosystem and evaluate strategies to mitigate corrupted information. Overall, the field of information integrity research requires much stronger research training and use of methods that ensure the credibility, reliability, reproducibility, and robustness of research findings that move beyond the exploratory phase. This would include, but not be limited to, advances in general-purpose measures, metrics, cross-platform instrumentation, cross-cultural methods, multimedia analysis tools, sampling techniques, and methods to assess statistical validity. Such advances are needed in the detection of corrupted information, context-sensitive characterizations, supply chain analysis, and in the assessment of harms, impacts, and outcomes.

Measuring the complexity of corrupted information. Research on the presence of corrupted information has mostly focused on detection, labeling, and deep fake forensics. For instance, researchers have made progress in methods for assessing individual pieces of content using techniques such as fact-checkers, key-word filters, and AI classifiers. However, work thus far also has significant limitations. A shortcoming in today's approaches has been bivariate scales that force complex phenomena into true/false or good/bad categorizations as opposed to more continuous measures of veracity, risk, or harm. Such framing lacks nuance around genuine uncertainties as well as the relationship between information and people's existing knowledge, beliefs, and values. Today's approaches also tend to focus on text versus other media, and detection capabilities often fail when applied to emerging content distribution technologies or platforms. In addition, detectors sometimes fail when corrupted messages are rephrased or reworded, causing machines to overlook variations that people intuitively recognize as having the same meaning. Therefore, advances in detection and instrumentation are needed to establish a multidimensional approach to measuring information integrity across platforms, media, and emerging technologies.

Context-sensitive characterizations. Beyond mere detection, the field requires better measures of the dimensions or factors that make a message or campaign or communication false relative to available evidence; for example, a demonstrably false claim versus a partly unsubstantiated claim or an accurate statement used to imply a faulty conclusion. Research is needed to address time-varying properties, such as an emergency warning that is true when issued but becomes false or misleading as time passes. Just as critical are the measurement challenges arising from changes in context: a false claim may include not only an assertion itself, but also a false narrative, coded language, misleading evidence, faked images or video, or logically fallacious arguments. Actors who create corrupted information also sometimes co-opt higher-integrity information by using selective quoting, misrepresenting context, and framing materials to interfere with people's reasoning or manipulate belief structures. Furthermore, manipulated information is sometimes copied into a context where someone seeks to provide a counter-argument, so the overall message might make high-integrity points even through part of the narrative contains manipulated information. Therefore, information integrity research should make advances in context-sensitive analysis for better detection and understanding of corrupted information.

Supply chain analysis. Robust measures of supply chains of corrupted information, including its sources, multilingual transformations, mechanisms of spread, and destinations, are also needed. To date, actors might be characterized by their association with corrupted information artifacts identified by independent fact-checkers or their divergence from information sources that follow rigorous journalistic standards. There has also been work on analyzing information campaigns, through modeling and detailed analyses of individual campaigns, and on understanding how spam markets work.[49] Meanwhile, the process by which corrupted artifacts are created—often in non-mainstream platforms—is generally less visible, creating a need for inference techniques that can expose similarities among manipulated artifacts as they move through a supply chain. The limitations reinforce that advances are needed in independent, cross-validated measurement of supply chains, recipients, and sources across highly complex, dynamic information ecosystems.

Assessment of harms, impacts, and outcomes. Advances also are needed in measuring the harms and impacts posed by corrupted information, starting with choosing and operationalizing fundamental conceptions of harm. Harm assessment has been complicated by differing public perceptions and priorities as to what is harmful; by fast-moving events such as disasters, in which the harm may be immediate but widespread and hard to document; and by the existence of harm that is hidden from public view, such as suicide attempts, misuses of medication, bullied children, and some threats of violence. Given the many types, levels, and operationalizations of harm, there is also a major measurement gap for rigorous evaluation of proposed strategies to reduce harms related to corrupted information.[50] Despite many purported remedies, many have not been replicated or "work" only for a subset of people, or even have "backfire" effects. Thus, gains in measurement science, along with parallel work on ethics guidelines, are also needed to accelerate the evaluation of the success and failure of proposed strategies.

Greater attention to the rigor of methods is required in research intended to causally connect harm to corrupted information sources, strategies, and recipients. This work will also require in-depth qualitative research for insights on people's perceptions of communications and harm in the context of their communications and lives. As argued in a recent analysis of measurement of corrupted information in science,[51] the operationalization of corrupted information needs to be clear, along with steps to increase reliability and validity of measurement, and to ensure that measures are transparent and results are replicable.[52]

3.1.3.1 Research Actions: Advance the Science of Information Ecosystems Measurement

- Develop consistent, transparent, cross-platform measures, statistically valid sampling methods, and efficient institutional review board processes for real-world assessments.
- Develop multidisciplinary testbeds and testing frameworks of information corruption, harms, harm mitigation strategies, and impacts on people, groups, and organizations.
- Make data more broadly available for analysis; create open-source repositories, training programs, researcher networks, and other mechanisms for researchers to learn; and share advanced methods of measurement and analysis, ranging from cross-platform instrumentation to survey design.
- Develop reusable methods to detect and characterize the full spectrum of corrupted information; its sources and destinations; and its interactions with beliefs, available evidence, and context.
- Advance approaches used to measure and assess causality, especially in relation to harms and harm-mitigation strategies.

3.1.4 Measuring Progress in Priority Area: Analyze, Model, and Measure Information Ecosystems

- Increase the number of multidisciplinary models that have undergone rigorous evaluation and make empirical connections between the human elements, ecosystem dynamics, integrity properties, benefits, and harms.
- Increase the number of open, shared, at-scale datasets for comparative analysis.
- Broaden and strengthen the reusable multidisciplinary toolbox of measurement techniques to assess causality, conduct outcomes research, and investigate connections between online behavior and real-world actions.
- Enable theoretically sound impact evaluation for a larger number of incentive structures, harm-mitigation strategies, events, and threat actions.

3.2 Priority Area: Investigate Safeguards that Assist People

People, communities, companies, and organizations who encounter corrupted information or work to mitigate its effects would benefit from effective strategies to foster resilience. Given that corrupted information is easy to produce, sometimes becomes free to distribute on a wide scale, and cannot be easily stopped at its source, a variety of safeguards should be considered. This priority area focuses on safeguards that reduce people's vulnerability to corrupted information.

There is no one-size-fits-all approach to designing and evaluating safeguards. The effectiveness of a particular method depends on the type of mechanism, when it is deployed, by what method, by whom, and how unique characteristics of the target population shape their experience of it. Necessarily, these dimensions interact with one another; they also should be informed by fundamental investigations of the human elements at the center of the system. Well-designed safeguards must address psychological and sociological factors as well as the cognitive, emotional, and behavioral effects of corrupted information on content recipients. They must also respect the autonomy of participants in the information ecosystem, safeguarding free expression and helping people make informed but free assessments of information.

The research objectives regarding safeguards that could reduce people's vulnerability to information manipulation are informed by four overarching interests: (1) deploying appropriate metrics for assessing effectiveness, (2) understanding the effectiveness of current approaches using rigorous methods, (3) improving the effectiveness of good approaches thus identified, and (4) developing new approaches. This priority area advocates an approach that keeps these overarching interests at the center of R&D efforts related to particular safeguarding strategies.

3.2.1 Research Objective: Develop Safeguards for Specific Instances of Corrupted Information

One class of individual safeguards aims to reduce the harmful effects of specific instances of information corruption by countering it with rebuttals, framings, or narratives that help people critically consider the corrupted information. Countering low-integrity information with high-integrity information is a well-established practice (sometimes referred to as counterspeech) that avoids censorship and enables free expression. Counterspeech can be used both proactively and reactively.

Proactive approaches to safeguarding individuals. A potentially promising approach, but one that may require prior knowledge of corrupted information that recipients might encounter, is inoculation or "prebunking." Inoculation is the technique of sharing a limited version of impending false information and either warning of its inaccuracy or exposing its weaknesses. This technique has been tried in such contexts as climate change and violent extremism with some encouraging results,[53] though with unknown longevity or effectiveness in resistant or hard-to-reach populations.[54]

Reactive approaches to safeguarding individuals. In the absence of foreknowledge related to corrupted information, more reactive methods may be necessary. Researchers and practitioners tend to call counterspeech focused on showing the falsity of certain information debunking. Debunking can take a variety of forms, including fact-checking—refuting incorrect statements with accurate ones, or annotating content with truth labels (i.e., "content labeling")—and counter-narratives, or countering false narratives with different, more thorough explanations of underlying facts. Research has shown that both fact-checking and counter-narratives can serve as effective safeguards under certain conditions. Fact-checking often works best when the underlying domain being discussed is relatively well understood and information is easily verified. Counter-narratives may be more effective in situations where facts are missing or there is lack of clarity about the underlying domain of discussion.

Yet, research on counterspeech strategies is still in its early stages. It remains unclear what types of information and counter-narratives are most compelling for which populations, in which subject area domains, and under what precise conditions. How, when, and how often to present these safeguards in the context of the platforms where people consume information are also open questions. It is also unclear whether counter-narratives or fact-checking lead to any positive effects that endure over time, or how the choice of channels impacts the effectiveness or durability of these approaches. Finally, further research is required to understand what metrics would be necessary to measure the effectiveness of the safeguards.

3.2.1.1 Research Actions: Develop Safeguards for Specific Instances of Corrupted Information

- Develop methodologies to assess the causal chain between information manipulation techniques, people's absorption of the information, and potential countermeasures such as inoculation, counter-narratives, and fact-checking.
- Develop participatory models wherein impacted communities are directly engaged in creating and deploying culturally and linguistically relevant countermeasures.
- Identify factors that affect the practicality or effectiveness of safeguards, including, but not limited to, platform design elements, community participation, characteristics of individuals or organizations, information sources, or threat actors.
- Disaggregate studies of effectiveness by demographic and other characteristics to understand potential variation and any disparate impacts.
- Advance natural language processing, information retrieval, and related techniques that can rapidly help people construct useful debunking and counter-narrative content in the face of novel manipulated information, and that can personalize that content to best meet a particular individual's needs for assessing information.

3.2.2 Research Objective: Assess How Information Sources and Cultural Contexts Impact the Effectiveness of Safeguards

People receive information from multiple, diverse sources. Research indicates that trust in the information provider is a key factor determining whether an individual is receptive to information others provide.

Understanding the network of social relationships in digital space that enables certain kinds of information to reach certain individuals is important to building effective safeguards. This includes (1) how individuals' relationships in offline spaces impact the channels through which they receive information online, (2) the extent to which various characteristics of individuals' social networks (online and offline) predispose them to consume or believe manipulated information, and (3) and how these same relationships—and the related information flows in online spaces—impact the effectiveness of various safeguards.

Information manipulation is enabled by and uses cultural, historical, and other societal contexts and opportunities. Likewise, safeguards against information manipulation effects must be grounded in the cultural, historical, and other societal contexts to be effective.[55]

3.2.2.1 Research Actions: Assess How Information Sources and Cultural Contexts Impact the Effectiveness of Safeguards

- Examine the relationships between social networks, information flows, and safeguard effectiveness.

- Determine what makes different kinds of sources more or less trusted by an individual, and how these patterns may differ based on the type of information being shared.
- Analyze the individual- and network-level factors that make information providers effective in reducing the negative impacts of manipulated information.
- Identify cultural factors and assess their influence on persuasiveness of information manipulation, people's susceptibility to information manipulation, and people's resilience to information manipulation.

3.2.3 Research Objective: Enhance Education on Media, Digital, and Information Literacy

Education and information literacy (to include scientific, health, media, and digital literacy) are important areas of consideration in this Roadmap. Education or training could potentially build resilience in children and adults to manipulated information; help offer tools to create, consume, and propagate trustworthy information; and teach them how to identify low-integrity information. A well-informed citizenry that is knowledgeable about the safe use of digital technology and media could be highly effective in combating information manipulation.

Educational safeguards, such as media, digital, and information literacy efforts, aim to foster cognitive tools and knowledge that help people consider information critically. These include teaching individuals to identify reliable sources of information, seek out balanced content and independent verification of claims, and identify hidden narratives in messaging presented to them. New strategies are being investigated using a range of evaluation approaches, including high-quality randomized controlled experiments, to help people reject false information, even to resist attempts at radicalization and violent extremism. Going beyond the content-focused safeguarding techniques described in Section 3.2.1, educational safeguards represent a longer-term approach to addressing the negative impacts of manipulated information.

There are several gaps in current educational, digital, media, and information literacy strategies and pathways that need to be addressed. First, research is needed on information literacy support mechanisms outside of schools and other formal learning environments. Second, both formal and informal support mechanisms need to be designed with inclusion in mind, aiming to reach marginalized and at-risk populations. Third, research is needed on how to make these support mechanisms more adaptive to individual and social needs, as well as to take advantage of evolving technologies such as conversational agents and virtual reality. Fourth, outcomes research is needed to evaluate efficacy across demographics and contexts. Finally, education and training need to be considered as an integral part of the development of platforms or other communication technologies, rather than as secondary or post-deployment concerns.

3.2.3.1 Research Actions: Enhance Education on Media, Digital, and Information Literacy

- Conduct longitudinal, cross-cultural studies of what allows young people to cultivate a critical thinking mindset early on, giving students practical tools to detect and successfully handle real-life information manipulation situations.
- Explore what types of educational strategies might increase resilience across diverse information manipulation tactics and topics (e.g., public health or consumer protections).
- Leverage emerging interaction techniques to develop more engaging and effective educational support mechanisms.

- Understand what enables successful life-long learning of digital literacy, media literacy, information literacy, and critical thinking skills outside of formal educational settings (such as through libraries and museums), such that people from different age groups, demographics, digital literacy levels, and cultural/language backgrounds might benefit from accessing this informal learning.
- Involve education researchers, librarians, and museum professionals centrally in this effort to identify appropriate support mechanisms and aid in empowering the user to make informed choices.

3.2.4 Measuring Progress in Priority Area: Investigate Safeguards that Assist People

- Evaluate a growing number of safeguards, strengthen the rigor with which they are evaluated, and use empirical findings to improve their effectiveness.
- Increase the accuracy with which an effective safeguard can be selected for a particular individual, population, or scenario.
- Increase the number of contexts in which safeguard failure is understood and accurately predicted.

3.3 Priority Area: Envision Technical Approaches to Enhance High-Integrity Information Exchange

While safeguards can assist people who encounter corrupted information, ecosystem design influences the way in which communities interact with high- and low-integrity information. Design decisions within information ecosystems create intended and unintended effects that modulate the creation, prevalence, and persistence of corrupted information. User actions affect personalized recommendation algorithms and amplification mechanisms to filter information, influence what kinds of information is spread, to whom, and how quickly.[56] This, in turn, can influence how people engage with the platforms themselves and what information they choose to create, leading to complex trade-offs between content creators' and readers' interests, information integrity, and corporate business models.

Design choices about interfaces, algorithms, interaction mechanisms, and system policies may affect these tradeoffs. However, research on the causal relationships between ecosystem design and information integrity remains in its infancy. Even straightforward design aspects, such as featuring items that early adopters interact with, can have profound and chaotic effects on what becomes popular,[57] with effects on emergent phenomena such as polarization, divisiveness, and outrage.[58] Meanwhile, tweaks to message passing mechanisms, such as limiting the number of times content can be reshared,[59,60] or to algorithms, such as reducing the amplification of content that receives angry comments,[61] can significantly affect the type and velocity of information moving through platforms. Furthermore, online experiences, such as anonymously chatting with someone who has different views, have been shown to reduce polarization,[62] which could inform design of new platform features.

Ultimately, fundamental research on ecosystem design must uncover mechanisms and causal relationships between design choices, societal benefits, and harms related to corrupted information. This section explores technical and sociotechnical approaches to designing next-generation information ecosystems that make it easier for communities to preserve information integrity. These R&D objectives investigate the emergent risks and benefits of design choices, explore mechanisms to help online communities engage in constructive civic discourse, evaluate technologies to help people establish trustworthy information channels, and assess how user customization might influence the integrity properties of information ecosystems. Collectively, this research should assess the feasibility of next-generation technologies, guard against unintended consequences, and de-risk prototypes to accelerate private-sector adoption of socially responsible technologies.

3.3.1 Research Objective: Develop Approaches for Socially Responsible Innovation

Being able to assess the effects of algorithmic, interface design, and interaction mechanism choices is necessary, but not sufficient, for improving the design of information platforms. Consideration of these effects, in terms of ethical, legal, social, and other implications, needs to be incorporated into the design methods and objectives used by information platform companies.

Attention to social implications is not a new concept, dating back to the 1930s with the development, study, and practice of corporate social responsibility.[63] Since then, design methods that focus on centering users, considering the needs of multiple communities, and addressing values have been developed across a range of design domains with the goal of encouraging more responsible innovation.[64] However, these methods are often not used in practice. Many designers and organizations are unaware of such methods, and those that are aware might be more motivated by immediate concerns with time to market or cost control. Further, envisioning possible negative outcomes of designs is difficult and has limited tool and method support. Nonetheless, the field of cybersecurity, which historically encounters similar tensions, continues to strengthen tools to incorporate protections at design time, improve their cost effectiveness, and measure progress toward security objectives. Similar advances are needed for information integrity, with the caveat that integrity researchers must work with users and communities to develop appropriate approaches.

Design of recommendation, amplification, and ad placement algorithms warrants special attention in terms of characterizing their impacts and investigating design choices that reduce aggregate levels of corrupted information. Advances in socially-responsible outcomes metrics, algorithmic monitoring, and algorithmic accountability tools are needed to quantify how algorithms affect relative exposure to accurate or inaccurate material across user demographics, patterns of interaction, and varieties of fraudulent content, as well as ways to evaluate whether platforms are meeting both business and wider social objectives. Consideration of malicious actors, as is increasingly common in the expanding science of adversarial machine learning and in applications of machine learning, is also important in information ecosystems. Assessment of designs' impacts on information integrity and on harms is also critical.

Finally, holistic outcome monitoring is needed to evaluate which combination of platform protections and human-centered safeguards are most effective. For instance, comparative studies would elucidate interactions between educational content, inoculation techniques, algorithmic safeguards, content labeling, and other safeguards over short and long timeframes.

3.3.1.1 Research Actions: Develop Practices for Socially Responsible Innovation

- Develop theories and formalize systematic approaches to explicitly integrate ethical and social considerations, engage with communities (e.g., community advisory boards); and leverage ethical, legal, and social implications panels throughout the research and design process for information ecosystem technologies, including media, search, marketing, and private messaging platforms.
- Increase the utility and uptake of design methods that measure social implications and potential negative impacts, including better ways to anticipate, monitor, and mitigate unintended outcomes as well as techniques to assess attack vulnerability.
- Advance research into societal outcomes measures, algorithmic monitoring, and algorithmic accountability tools, to include fundamental research in multi-objective optimization and related areas that allows for flexible expression and satisfaction of simultaneous goals.

- Design accountability mechanisms to support both internal and external auditing to assess whether information platforms are meeting these objectives during both testing and deployment.
- Advance computer science, data science, and information systems curricula to increase students' understanding of their social and ethical responsibilities with regard to information integrity.

3.3.2 Research Objective: Innovate to Support Constructive Engagement Around Information

Beyond the integrity of information itself, questions arise around support for high-integrity open dialogue and civic discourse around that information, two elements essential to a well-functioning democracy that values free speech. Empirically, bad actors can exploit today's platforms to disseminate corrupted information and steer conversations to intensify societal divisions,[65] but less is known about how technologies can make it easier for communities to nurture constructive discourse and move toward high-integrity conclusions.[66]

Promising objectives include designs that encourage meaningful dialogue between those with different views. Future research might explore designs that amplify common ground above polarizing content, help people thoughtfully engage with information in pursuit of high-integrity conclusions, maximize the number of people who feel included and heard in civic discussions, and investigate troll-resistant chat assistance to reduce hostility.

3.3.2.1 Research Actions: Innovate to Support Constructive Engagement Around Information

- Design platforms and algorithms that help communities foster productive discourse, especially for topic areas characterized by uncertainty, complex risk-benefit trade-offs, or lack of ground truth.
- Develop designs that incentivize prosocial engagement across polarized groups, make it easier for communities to engage in constructive dialogue about ideas, and assist communities in respectfully mitigating disruptive behaviors.
- Design interfaces that help individuals, communities, and crowds assess information integrity in the context of their own needs and norms, supported by computational analysis methods and metrics described in other parts of this Roadmap.
- Explore how ecosystem design can help communities normalize high standards of information integrity by cultivating social conventions around ideas and dialogue, and by providing tools to support those conventions.

3.3.3 Research Objective: Design Trustworthy Information Channels

Gaps also exist in platform technologies and sociotechnical approaches to construct, recognize, and maintain trusted information channels in an increasingly digital age in domains that include news, science, and public health. Building, maintaining, and, when necessary, rebuilding trust in channels that strive to provide high-integrity information are key challenges in the face of manipulated-information campaigns designed to damage that trust.

There are several approaches toward building this trust, many of which revolve around transparency and attribution. Promoting journalistic and scientific standards in the production of information and providing ways to assess whether processes and outputs meet those standards, is one path toward achieving trustworthy information channels. Tools for establishing the provenance of information, such as content authentication, digital watermarking, and digital signatures can help address problems around synthesized or deliberately corrupted information.

Addressing uncertainty and complexity is another important aspect of building trust. Entities that create or distribute information could indicate incomplete, uncertain, contested, and evolving aspects of information to support appropriate caution around over trusting high-integrity but still-evolving information, as well as guard against attempts to exploit or artificially create those uncertainties. New methods for translating complicated information to multiple audiences in a high-integrity manner are also needed.

Finally, there is value in high-integrity channels that extend beyond single sources or platforms to interconnected information ecosystems. Even platforms and content sources that strive for high integrity pose risks if they become single points of failure or are easy to circumvent. In the absence of gatekeepers, this suggests research into approaches that maintain integrity and trust across decentralized media, through either crowd-based methods or other decentralized means to preserve integrity indicators and distribute, attribute, and assess information across multiple sources, platforms, and ecosystems.

User customization. Standardized protocols and new architectural models could create a novel class of interface technologies that enable users to customize recommendation algorithms, content moderation preferences, and other media choices. Greater user customization could be enabled through improved algorithmic transparency or through a new class of technologies that sits between media platforms and the user. Notionally, this new class of technologies would aggregate content from user-selected platforms, enable users to adjust algorithmic preferences, and give users the option to select curators they trust to maintain information integrity standards on their behalf. Nonetheless, research is needed to explore how user customization would affect financial incentives for content creation, viable business models, and the prevalence of high-integrity information within information ecosystems. Furthermore, research on data interoperability, algorithmic transparency, and sociotechnical processes would be key to explore feasibility and the ability for novel architectures and standards to enable trustworthy information channels.

Overall, comparative evaluation is needed to determine which design approaches would result in trustworthy information channels, their limitations, and what protections must be built in to guard against unintended consequences.

3.3.3.1 Research Actions: Design Trustworthy Information Channels

- Design information channels with built-in source verification, provenance tracking, and integrity indicators.
- Investigate how increasing social polarization affects people's assessment of trustworthiness and what properties information channels must have to be trusted by a wide variety of people.
- Develop tools around information integrity for rigorous, collaborative auditing of platform and source practices with transparency, provenance, and explanation capabilities.
- Create methods to represent legitimate uncertainties in high-integrity information, detect attempts to create unwarranted uncertainty, and communicate knowledge in ways appropriate to user groups, use cases, and contexts.
- Investigate the properties of open, decentralized information ecosystems and develop integrity protections, including novel architectures, proposed standards, and application programming interfaces, to enable user customization and cross-platform and cross-participant safeguards.

3.3.4 Measuring Progress in Priority Area: Envision Technical Approaches to Enhance High-Integrity Information Exchange

- Evaluate, refine, and increase the number of measures, outcomes metrics, algorithmic monitoring, and algorithmic accountability tools to assess causal effects between design elements and socially responsible outcomes.
- Evaluate a growing number of technologies to enable productive engagement around information and establishment of trustworthy information channels.
- Evaluate indicators and warning mechanisms to assess the health of an information ecosystem and detect when it may be at risk for, or under attack, by manipulation.
- Create tools that move beyond measuring social-responsibility properties toward proactive design of systems that make it easier for communities to cultivate high-integrity outcomes.

3.4 Priority Area: Understand Effective Strategies to Address Manipulated-Information Campaigns

Campaigns to spread corrupted information deserve special attention because of their dynamism and their continuous development of new tactics to circumvent detection and measurement. The expansion of global digital connectivity, immersive and intrusive information technology, and widely accessible digital marketing techniques is enabling rapid growth of a corrupted information industry. The "disinformation-for-hire" global market has become a "boom industry," outpacing the ability of governments to keep up.[67] Harmful information campaigns deceive audiences by integrating corrupted information into narratives over months or years. What distinguishes these campaigns from other types of mis- or disinformation is the methodical use of corrupted information to change the attitudes, beliefs, and actions of a cultivated audience over time. This section explores sociotechnical mechanisms to address information campaigns that result in harm such as negative health outcomes, financial theft, radicalization, undermining international alliances, and degradation of democratic processes.

Campaigns initiated by foreign states, criminal organizations, and other malicious actors are of particular interest because they can impact national security and the security and safety of individuals, organizations, and communities. Given that campaigns may originate anywhere and cross international borders, researchers must trace them across jurisdictions to enable comprehensive analysis. Research requires a multimodal perspective as campaigns move across numerous online platforms, television, print media, book reviews, and offline channels. Further complicating matters, platform policies and investment in safeguard technologies vary across geographies, so domestic research may fail to generalize across international borders. Thus, analyses of regional, multilingual, and cross-cultural variations are needed, especially in areas of the world where research investments are sparse.[68] Finally, interdisciplinary approaches remain essential, because research cuts across diverse fields including AI, national security, political science, education, marketing, communications, entertainment, behavioral science, data science, natural language processing, and expertise domains associated with populations targeted by manipulated information.

Campaign de-escalation tools require continuous feedback loops to detect harmful activities, develop appropriate responses, and measure the impacts of any actions taken. The following subsections describe techniques to anticipate, de-escalate, and respond to campaigns that cause inadvertent harm and rapidly adapt to tactics associated with foreign state manipulations or criminal activity. Section 3.4.4 describes assessment techniques to converge on effective response strategies.

3.4.1 Research Objective: Anticipate Harmful Manipulated-Information Campaigns[69]

Campaign tracking requires accurate, scalable techniques to identify and anticipate manipulated narratives, expose falsification technologies, identify perpetrators and tactics, and monitor campaign evolution. Capabilities to distinguish harmful campaign content from harmless material are needed for an increasing diversity of campaign topics, contextual variations, languages, actors, and audiences. Research must address non-textual and multimodal, multilingual, cross-platform campaign content that exploits gaps in today's detection capabilities as well as campaigns on emerging platforms for which practitioners lack threat frameworks, comprehensive detection, and measurement tools.

Technological advances must accelerate adaptation to an ever-growing set of platforms and campaign obfuscation techniques such as subtle narrative manipulations, multimodal deep fakes, and ever-changing code words. Research should go beyond static content detection to address dynamism, as actors deliberately adapt campaigns based on feedback and as campaign messaging evolves organically. Because actors may employ automated bots, troll farms, or unsuspecting bystanders to amplify their influence, research is needed to expose digital fingerprints for technologies that groom audiences, falsify content, and inauthentically amplify campaign material. Further, research is needed on how attribution or provenance-tracking technologies affect the ways influence campaigns generate and distribute manipulated information, and whether those changes may support earlier identification of these campaigns.

Ultimately, capabilities need to move beyond detection to provide early warning systems for the manipulated-information campaigns that are most likely to become influential. Proactive communication strategies designed to foster resilience to manipulated narratives would greatly benefit from early warning capabilities to enhance the flexibility of planning and de-escalation efforts. Forecasting requires research into where manipulated-information campaigns tend to emerge to monitor new narrative formation and expose tactics for audience engagement, persuasion, or polarization. Research on predictive tools would enable responders to anticipate where and when campaigns might emerge, particularly around breaking world events, and enable principled studies of campaign spread, mitigation strategies, and population response.

3.4.1.1 Research Actions: Anticipate Harmful Manipulated-Information Campaigns

- Improve detection, tracking, and forecasting of harmful campaign narratives and content involving manipulated information, including their evolution over time and across ecosystems.
- Develop forensic and provenance-based techniques for detecting information manipulation across media formats and platforms, including videos, audio-only chat room dialogues, crowd-sourcing platforms, consumer sites, massive online multiplayer games, and emergent platforms, such as extended reality.
- Research and develop technologies that support the identification of automated bot accounts, deep fakes, and other falsification technologies.
- Accelerate detection of nascent information campaigns and adaptation to new attack vectors, evolving campaign content, shifting campaign tactics, and novel falsification technologies.
- Scale tracking capabilities to the size, complexity, speed, and diversity of modern manipulated-information campaigns, while exceeding the accuracy of manual alternatives.

3.4.2 Research Objective: Develop Proactive and Effective Strategies to De-escalate Manipulated-Information Campaigns

Addressing manipulated-information campaigns requires a multipronged approach that both preserves trust with target populations and deters nefarious actors. This section explores ways to meaningfully engage with target populations. Responders, such as those in public health, emergency preparedness, consumer protections, law enforcement, disaster response, or community leadership, would benefit from best practices to mitigate or de-escalate manipulated-information campaigns that threaten to harm a constituent population or the integrity of their operations. Next-generation techniques must go beyond traditional communication strategies. Instead, tools are needed to alert responders of previral campaign narratives that might cause harm and enable the responders to fill information voids before corrupted information arrives. Understanding a population's susceptibility to influence and the potential severity of negative consequences can inform efforts to preempt campaigns through prioritization, education, inoculation, and other tactics. For example, a promising line of work examines better ways to communicate evidence and facts through easily comprehended visualizations and "gist" communications connected to people's core values and trusted groups.[70] New sociotechnical approaches are needed to identify effective inoculation techniques, tailor content to audiences, and advise on communication strategies that foster resilience.

As an example, the Census Bureau took a proactive approach to anticipate and address false claims about the 2020 Census, using inoculation and a rapid-response multistakeholder team.[71] They worked with a vast, nationwide network of trusted messengers and community-based organizations to communicate with hard-to-reach populations. Their efforts predicted and diffused campaigns that aimed to suppress census participation in select demographics or sought to undermine trust in the reliability of the census.

Research is needed on high-integrity strategies to respond to concerns and emotionally sensitive issues, to develop effective campaign response tools that inform the audience, listen to their perspectives, and retain and build trust. Response tools could assist with risk communications, alleviate fears about new technologies or health interventions, and help scientific researchers design studies that address areas of public concern. Research is needed to develop evidence-based strategies that foster bidirectional dialogue, build or rebuild trust, address areas of public concern, and inoculate audiences to harmful information campaigns. Finally, response tools must keep pace with the scale and speed of modern information campaigns to mitigate harm. The volume, velocity, and variety of messaging requires significant levels of automation, aided by advances in AI techniques to accelerate and scale evidence-based strategies.

3.4.2.1 Research Actions: Develop Proactive and Effective Strategies to De-escalate Campaigns

- Develop tools to help campaign de-escalation teams prebunk corrupted narratives before they gain traction.
- Integrate campaign monitoring with curriculum development, scientific research design, educational research, communications, and outcomes research activities to inform next-generation safeguards and campaign de-escalation techniques.
- Develop evidence-based tools, strategies, and tactics to respond to viral campaigns in a manner that builds trust and resilience to harmful corrupted information.

- Develop evidence-based strategies, community engagement strategies, and participatory practices to build trust and confidence in government institutions.
- Develop sociotechnical techniques and automation to accelerate response, strengthen high-integrity communications, and establish trust across diverse populations and influencers.

3.4.3 Research Objective: Develop Deterrence Techniques for Foreign State Manipulations and Illegal Activity

Within a national security context, next-generation capabilities are needed to make it harder for foreign state competitors and adversaries, extremist groups, and criminal organizations to conduct manipulated-information campaigns. Today's defensive techniques tend to be reactionary and may respond to each new campaign message in a piecemeal fashion. In contrast, future capabilities should enable responders to disrupt an adversary's campaign in the early stages, slow their operations, increase their costs, and reduce their influence in a comprehensive way. Such capabilities should integrate diverse tactics, including, but not limited to, disrupting falsification and amplification technologies; distracting with neutral content; reducing access to target populations; interrupting corrupted information supply chains; delaying transmission of campaign content; and coordinating parallel actions via diplomatic, military, and economic channels. Defensive techniques should also include strategies to deter adversaries from manipulation, mitigate harm, and increase the societal resilience of the target audience.

The United States might have limited time (hours or minutes) to detect and respond to adversary activity before it starts to have a harmful impact. This requires timely assessment of what to address, rapid decision-making about how to respond, acquisition of the necessary legal and policy approvals, collection of resources for actions, and execution of actions. Defensive capabilities would thus benefit from timely, accurate, continuous feedback regarding both the status of the adversary's campaign and the effectiveness of response activities so that strategies can be dynamically tailored at the pace of relevance.

Campaign disruption may require attribution of nefarious actors, foreign actors posing as Americans, cyber tactics, and campaign infrastructure. Multiplatform attribution continues to present challenges, because threat actor detection technologies need to go beyond surface profile and content features to detect coordinated accounts on multiple platforms over time. Threat actors also act through proxy actors such as troll farms for hire. Research should explore the viability of privacy-enhancing technologies to track illicit behaviors and troll farm activity across platforms while protecting the privacy of legitimate users. Techniques are needed to enhance collaboration across entities, sectors, and jurisdictions to help tackle shared challenges such as financial crime, use of corrupted information to deliver cyber payloads, human trafficking, and health crises.

3.4.3.1 Research Actions: Develop Deterrence Techniques for Foreign State Manipulations and Illegal Activity

- Develop tools, technologies, and signatures to attribute adversary campaigns and disrupt corrupted information supply chains such as those associated with private adversaries residing in Russia.
- Create and validate forensic tools to infer foreign state ownership or control of media outlets that spread manipulated information.
- Elucidate adversary coordination across multiple platforms, channels, actors, and gray zone activities such as military exercises, troop movements, or economic coercion.

- Develop tools and techniques to assess and quantify risks associated with different information channels.
- Identify new malign activity signatures and develop complementary analytics, machine learning, and privacy-enhancing technologies that enable detection of malicious influence and illegal activities as well as tracking of nefarious actors, troll farms, and associated automated bots across platforms without compromising user data.

3.4.4 Research Objective: Enable Outcomes Monitoring

Responders would benefit from timely, accurate, quantitative and qualitative impact assessments of harmful information campaigns and real-time approaches to de-escalate the spread of manipulated information. This requires moving beyond traditional, but often problematic, metrics of online media engagement, such as views, likes, and shares, to a broader array of online and real-world metrics customized to the campaign and target audience. Research should explore both offline and online indicators to measure meaningful changes in behavior, ranging from real-world actions—such as protests, tipline calls, or symbol displays—to changes in a target audience's information consumption patterns, language, or patterns of community engagement. Aggregate measures of effectiveness, ideally evaluated in collaboration with platforms, are also needed to investigate whether approaches reduce the incidence of fraud, harmfully corrupted information, criminal activity, and foreign state-sponsored manipulated-information campaigns.

Outcome-monitoring research should move beyond intermittent, static assessments toward reliable, real-time measures of effectiveness. Research is needed to elucidate the interplay between culture, belief structure, relevant events, and population response to both harmful information campaigns and strategies to de-escalate manipulated-information campaigns. Impact assessments are also needed to evaluate how well techniques address or navigate around systemic factors that incentivize and facilitate the creation of corrupted information.

3.4.4.1 Research Actions: Enable Outcomes Monitoring

- Develop timely, accurate, quantitative metrics and qualitative methods to monitor the impact of harmful information campaigns and response strategies, with an emphasis on behavioral effects relative to the prevalence of corrupt information.
- Collect and validate baselines along with outcomes measurements to enable accurate, reliable impact assessments.
- Continuously evaluate the effectiveness of campaign mitigation and educational strategies.

3.4.5 Measure Progress in Priority Area: Understand Effective Strategies to Address Manipulated-Information Campaigns

- Advance capabilities that reduce the human effort required to anticipate, track, and de-escalate manipulated-information campaigns.
- Evaluate and refine a growing number of approaches to increase the cost and difficulty for foreign states and other nefarious actors to perpetrate manipulated-information campaigns.
- Increase the accuracy and reliability of outcomes-monitoring capabilities for both de-escalation and deterrence activities.

4. Research Collaboratives and Infrastructure

Information integrity researchers face unique challenges when they strive to translate findings at the speed of relevance. Across translational efforts, outcomes research, or studies of internet-scale dissemination, some research problems are best addressed in partnerships that enable experts with complementary backgrounds to investigate novel approaches in real-world settings. The next two priorities highlight opportunities for joint investments across government, research communities, practitioners, and policymakers that would benefit from alignment of funds, talent, and resources. These priorities call for investments in multipurpose data and infrastructure and strategies to connect science with policy and practice.

4.1 Priority Area: Foster Data Access and Partnerships

Serious advances in information integrity research will require comprehensive, longitudinal data on the production, spread, digital routing, and consumption of manipulated information across multiple platforms and channels. Despite recent advances, researchers agree that many of today's studies are based on nonrepresentative samples of publicly available data. Lack of access to data both on how people consume and share information on platforms, and on how those platforms' algorithms and moderation policies affect what people see, poses a major barrier to fully understanding the integrity of the information exchange. This lack of platform transparency limits researchers' ability to understand the drivers of corrupted information, harmful group plans (e.g., hate crimes or violence), or associated cyberattacks.

Akin to impacts in domains such as public health or finance, research would benefit from consortia that allow many interested parties—platforms, researchers, regulators, and end users—to aggregate data, tools, methods, and insights in a coordinated, responsible, transparent way. The requisite research infrastructure would require large-scale sustained funding models. As in healthcare, tiered access models would provide flexibility to adjust data visibility based on the nature of the research being performed and protections required. Such research consortia would need to grapple with issues pertinent to information ecosystems, such as adequately protecting proprietary information, personally identifiable information, and other types of sensitive data, as well as the risk of deceptive narratives designed to undermine those engaged in research. Depending on the use case, data access conventions might vary across independent researchers, businesses, government practitioners, and funding agencies to ensure that partnerships respect relevant laws, authorities, and protections for statistical datasets.[72]

Ideally, data partnership strategies would aspire to streamline and publicize aggregate analyses in areas of broad public concern, while making analyses of sensitive data possible for independent researchers with appropriate expertise and approvals. Progress toward this vision will require a combination of access to data, improved data usability, partnerships, and enabling infrastructure to advance information integrity investigations. Data approaches developed in collaboration with businesses and by independent research consortia are both important, given the unique roles they play in promoting transparency and social responsibility.

Access to data. An understanding of the information integrity ecosystem, and threats to information integrity, rest on the availability of solid evidentiary data. When such data are not available, evidence may be misleading. Complicating the problem is that information ecosystems typically span multiple platforms; when researchers lack data from multiple platforms and must rely on one platform to have some public data or research-friendly policies, the conclusions they draw are likely to be skewed and limited. To understand the true reach and impact of corrupted information, researchers will need data from multiple media platforms and a variety of other information sources including TV, radio, blogs, and mobile apps.

Data usability. Improved metrics, privacy protections, cross-platform measurement tools, and validation methods are all needed to enhance data quality and usability. Techniques to align and normalize public datasets are also needed to address differences in data collection methods, sampling rates, or geographical aggregation. Comprehensive longitudinal datasets would benefit from measurement tools developed in partnership with platforms to enable researchers to define, scale, and validate new metrics across multiple information channels. Finally, additional research on privacy-preserving techniques, anonymization strategies, population measures, and tiered access strategies is needed to facilitate studies of information integrity effects while addressing the concerns about data and privacy protection described previously.

Partnerships. Comprehensive analysis of harm from corrupted information may require public-private partnerships akin to the data interoperability ones found in the health sphere.[73] As in other industries, researchers and corporate representatives could work together to identify mutually beneficial R&D goals such as addressing inaccurate information that results in health or safety harms to users. The Coalition for Content Provenance and Authenticity,[74] the Partnership for Countering Influence Operations,[75] and the research partnerships between businesses and the European Digital Media Observatory[76] may provide good models, as might platform efforts to develop Ad Libraries in accordance with the EU Code of Practice on Disinformation.[77] Ideally, such partnerships would enable consistent, transparent, cross-platform indicators such as exposure, prevalence, or velocity. Research partnerships to develop independent measurement techniques are important as well, especially in areas where information ecosystem entities have historically been reluctant to provide transparency. It is important to note that such public-private partnerships must be cognizant of the First Amendment and human rights and that government organizations should consult with counsel prior to establishing or engaging in these partnerships to ensure that their activities are consistent with applicable laws.

Enabling infrastructure. The research community can accelerate its efforts by investing in shared data infrastructure, engineering, measurement, analysis, and quality assurance resources. Data access sites, such as Harvard Dataverse and Kaggle™, have already proved useful. In addition to pooling data resources, data storage and computing resources could also be shared as these are often costly for researchers to acquire and maintain. Reusable software tools for data pedigree, provenance, and analysis would also accelerate research, especially if supported by vibrant software development communities and appropriate incentives for data owners, producers, and custodians to use them. Facilities to simplify common data preparation tasks and tools to collect or convert data into shared formats; capture metadata; clean, annotate, and analyze data; support comparative analyses; and host data-connected publications would enable researchers to efficiently build upon one another's work. As new perspectives emerge, new or improved annotations could add additional value to shared datasets.

4.1.1 Research Actions: Foster Data Access and Partnerships

- Create methods, whether through partnerships, voluntary codes of conduct, or legislation, for the provision of platform data at different levels of access, including public-facing dashboards with key summary information available to the public; aggregated data for reporters and non-governmental organizations, and fine-grained, large data for trained, vetted researchers (akin to trained data stewards in the Special Sworn Status program[78]).

- Develop privacy protections, measurement tools, and validation methods to maximize usability of data available for research.

- Develop research consortia in partnership with multiple corporate partners and work together to enhance measurement and technology R&D around mutually beneficial goals related to trustworthy information exchange and harm mitigation.
- Invest in common software and hardware infrastructure to enhance data collection, protection, analysis, and access.
- Develop shared tools for data pedigree, provenance, normalization, alignment, and analysis to accelerate research and support evaluation efforts.

4.1.2 Measuring Progress in Priority Area: Foster Data Access and Partnerships

- Establish and expand participation in information integrity research consortia.
- Converge on several shared use cases for outcomes research, technology R&D, and translational efforts.
- Work within consortia to develop and expand common application programming interfaces to scale outcomes research and enhance data access.

4.2 Priority Area: Connect Research to Policy and Practice

Ultimately, partnerships should move beyond data transparency and measurement toward improved outcomes. To achieve improved outcomes, effective collaborations will be necessary across fundamental, use-inspired, translational, and outcomes researchers, as well as practitioners and policymakers. Each contributor performs an important function, with fundamental researchers providing sound measurement techniques and generalizable knowledge, use-inspired researchers de-risking novel prototypes, translational researchers increasing the utility of those prototypes for policymakers and practitioners, and outcomes researchers measuring longitudinal effects in real-world contexts. Active collaboration among researchers, policymakers, and practitioners becomes paramount to provide practitioners and policymakers with the information needed to make informed decisions about technology and policy, and to ensure that researchers have the real-world context necessary to discover novel approaches that yield desirable effects.

Informing policy. In the case of information integrity, there are numerous domains where policy- or practice-based alternatives may help address harms associated with the spread and consumption of corrupted information. For example, policies can be developed to encourage thoughtful interactions between digital information and consumers, educate the American public on how to recognize and address corrupted information, and strengthen trust in democratic processes. Yet to build effective policy, sufficient empirical evidence (grounded in fundamental research agendas) is necessary on the topics targeted for policy action. Detailed empirical analyses, and the systematic knowledge of people and the world that they create, are imperative to developing effective policy- or practice-based strategies. Further, these insights can be used to develop tools based on game theory, agent-based modeling, and other approaches that might allow researchers and policymakers to explore a wide range of policy options or support mechanisms at low cost. For this reason, research on information integrity should aim to suggest concrete implications for policy as appropriate.

Actionable recommendations. Once adequate empirical information is available on key information integrity topics, and implications for policy and practice have been considered, there is still translational work to do. The gaps in priorities, incentives, and communication styles between researchers and policymakers are well documented, often leading important research to remain in the pages of journals rather than in the text of congressional legislation or Federal regulations. To address these gaps, concrete efforts are necessary among the research community to communicate the findings of empirical research clearly and concisely, with attention to how empirical findings may relate to potential alternatives in policy and practice.

Foundations for policy tools. A key pathway to evidence-based policymaking is the development and use of metrics to track the incidence of harms caused by manipulated information and to identify contexts in which evidence-based actions make a difference. For example, policy development and implementation should be informed by an account of the harm or harms that are attributable to manipulated or corrupted information. If possible, metrics should track the number of people affected, the scope or severity of the outcomes, and how the prevalence of corrupted information relates to the extent of the problem. Consistent methods to assess harms should be developed to evaluate and monitor the utility of actions taken to mitigate negative effects or build resilience. Additionally, developing metrics-based thresholds of corrupted or manipulated information activity will be necessary to assess what strategies are appropriate for a particular context and to aid in real-time selection of suitable alternatives. Thresholds for response should utilize empirical evaluations informed by past occurrences of potential or realized harm.

Prototype technologies. Likewise, pipelines are needed to translate empirical findings into prototype technologies that can be tested at small and large scale. Additional approaches, such as forming multisector project teams that work side by side to connect researchers with technology professionals (developers and practitioners), or researchers with end users, or researchers with policymakers, or people with all four backgrounds, also could form the foundation to test and refine research prototypes in realistic settings. These and other translational approaches can allow fundamental R&D to play a more central role in informing actionable, successful policy- and practice-based alternatives.

Incorporating evaluation of policies, practices, and technical protypes. It is critical to assess the impacts—potential and real—of policies, practices, and technologies developed to address corrupted information. Just as we must develop new empirical knowledge about topic domains related to information integrity, we must also develop knowledge about whether practices, policies, and technologies are effective and when they create unintended consequences. This aspect of the work is especially important for digital systems, wherein properties of the algorithms and interfaces that process and present information may lead to counterintuitive outcomes that change over time.[79] Taking this approach typically requires a formal research agenda, including the initial development of empirical questions related to the hypothesized efficacy of proposed actions. These questions should be posed at the very beginning of policy- or practice-based strategy development so that outcomes research on the policy, practice, or tool can take place alongside the application of the policy or practice. The positive feedback loop created by this approach will serve to strengthen the overall information ecosystem.

4.2.1 Research Actions: Connect Research to Policy and Practice

- Develop and strengthen translational pipelines that can assist citizens and researchers in continuously communicating fundamental R&D insights to appropriate policy and practice partners.
- Pursue efforts to translate the findings of fundamental R&D into actionable recommendations for decision makers and develop use-inspired tools to accelerate the effective application of research within real-world contexts.
- Embed formal evaluation of policies and practices into the initial development process, such that rigorous outcomes research to continuously track the impacts and implications of policies and practices is related to information integrity from the start.

- Include, as part of formal evaluation, an assessment of potential consequences and implications of policies and practices, whether intentional or unintentional, to help improve the efficacy of the policy- or practice-based alternatives.
- Include, once evidence of efficacy has been established, an assessment of public perception of the policies or proposed actions in the evaluation process.

4.2.2 Measuring Progress in Priority Area: Connect Research to Policy and Practice

- Increase the number of research prototypes and results that are adopted by policymakers and practitioners.
- Increase the percentage of research efforts in which policy or practice partners participate in research design and evaluation.
- Increase the percentage of translational efforts in which outcomes research collects longitudinal data, results are replicated, and evaluation of potential consequences occurs at the outset.

5. Action Plan

5.1 Roles and Responsibilities

This section identifies the roles for public and private organizations, and identifies strategies for ensuring coordination of information integrity R&D within and across these sectors.

5.1.1 Government Organizations

The coordinated R&D activities of this Roadmap are facilitated or carried out by Federal departments and agencies with varying missions but complementary roles. This arrangement ensures that the full spectrum of R&D approaches is represented and engaged. Each department or agency is encouraged to incorporate these research priorities into its research plans and programs, drawing on its individual strengths and in the context of its mission. Science agencies should solicit and fund multidisciplinary research, explore larger multi-year consortia that enable researchers to pool resources and investigate longitudinal trends, and always demand strong scientific methods in all funded initiatives. Service agencies should consider the results of R&D in meeting immediate and future mission requirements. Both science and service agencies should explore partnerships that advance translational and outcomes research, both within the government and between researchers, communities, and the private sector.

State and local governments play a key role in public health, elections, education, and community engagement. Collaborations between researchers and state or local governments can assist in evaluating information integrity approaches in local contexts across the country.

5.1.2 Academic and Research Organizations

The research community is critical in effectively addressing the research priorities in this Roadmap. Researchers are encouraged to consider both disciplinary and interdisciplinary approaches, measurable efficacy and efficiency, reproducibility of experiments, and strategies for transitioning successful research into practice when developing proposals and initiating research in this area. Leading scientific organizations can help advance this domain by convening experts from across the ecosystem to collaborate on developing solutions and providing guidance. Another important aspect highlighted in this Roadmap is the development of robust and effective education and information literacy pathways, both formal and informal, to facilitate a more informed and empowered citizenry.

5.1.3 Private Sector

Private-sector R&D is typically internal and focused on product-development goals based on the specific needs of the company as well as on profitability and turnaround time. Nonetheless, there are opportunities for the R&D activities of the private sector and the public sector to be synergistic and complementary. A robust information integrity research ecosystem will offer several mechanisms that can be mutually beneficial for the two sectors. This Roadmap outlines several opportunities for engagement with the private sector, but the most critical ones are transparent data access, common use cases for research purposes, and algorithmic transparency. Another partnership opportunity for commercial entities would be to jointly identify precompetitive research areas in which public-private partnership funding would be most productive.

5.1.4 International Partnerships

This Roadmap emphasizes the fact that information integrity is a global problem that necessitates a global response. There are several international organizations and entities working on synergistic activities in the information integrity space. Opportunities for fruitful collaboration exist in expanding efforts to engage with international partners to enable more collaborative research, cross-fertilization, and data sharing.

5.1.5 Individuals, Civil Society, and Community Organizations

Participatory research models should play a key role in information integrity research. As programs are developed, care should be taken to involve community participants and clearly communicate opportunities to participate in and influence the direction of research. Community engagement will be crucial to ensure that investigations of next-generation designs, safeguards, policies, and best practices are responsive to real-world problems, contexts, and public concerns.

5.1.6 Nonprofit, Philanthropic Organizations, Foundations, and Think Tank Groups

Various nonprofit organizations play an important role in building collaborations with entrepreneurs, businesses, researchers, government organizations, and communities. Information integrity research would benefit from participation by nonprofit organizations to build and maintain partnerships to facilitate data collection, data access, and evaluation of research for policy or community application.

5.1.7 Coordination and Collaboration

This Roadmap also calls for increased coordination and collaboration across all these sectors to make measurable and effective progress on this important problem and to avoid redundant research initiatives. It is important to convene meetings to jointly explore issues of mutual interest, share research results and best practices, develop standards for safeguard efficacy and evaluation, and explore new challenges. Federal agencies should consider public-private partnerships and new funding initiatives to cultivate a vibrant community of researchers working in this area. It is important to emphasize that even within public-private partnerships, private companies retain discretion and independence, consistent with applicable laws, regarding corporate activities that relate to information integrity.

Research coordination between Federal departments and agencies is facilitated by the National Science and Technology Council (NSTC). Unclassified Federal R&D efforts in networking and information technology are coordinated by the NITRD SC and its National Coordination Office. Classified research efforts are coordinated by the NSTC Special Cybersecurity Operations Research and Engineering Subcommittee. Each year, the NITRD Program compiles and produces a Supplement to the President's Budget Request that provides highlights of agency R&D activities in various areas of information technology and networking.

5.2 Concrete Actions

The Federal research agencies and the research community, in partnership with businesses, communities, nonprofit organizations, and other government institutions, could pursue the following actions to achieve the goals, priorities, and objectives outlined in this Roadmap.

5.2.1 Create a Firm Foundation for the Field of Information Integrity

Integration of behavioral and social disciplines with technical ones is paramount to tackling the sociotechnical challenges associated with the socially responsible innovation of information integrity. This requires rethinking the way that everything from funding mechanisms to curricula to academic research centers to public-private partnerships is structured. The following are several concrete actions toward this objective.

5.2.1.1 Reimagining Multidisciplinary Research for Information Integrity

- Foster a multidisciplinary (social, behavioral, economic, and computing and information sciences) research community around the science of information integrity by supporting research by teams that adopt a transdisciplinary approach.
- Enhance conferences, research consortia, and multidisciplinary research centers to advance the science of information integrity.
- Assess and strengthen funding models that support continuous development, curation, and maintenance of standardized data sets, common data enrichments, and cross-platform measurement capabilities and sustain those investments over years and decades.
- Expand common vocabularies, theoretical foundations, joint curricula, and new modalities of multidisciplinary collaboration to accelerate progress toward an information integrity field that is greater than the sum of its parts.
- Give preference to cooperative agreements and contracts that include multidisciplinary collaboration in pursuit of concrete objectives that are impossible for one discipline to accomplish alone (e.g., progress toward transdisciplinary theoretical foundations or use-inspired prototypes).
- Coordinate the social media and big data method training of undergraduate, graduate, and post-doctoral students in the social and computer sciences and related disciplines.

5.2.1.2 Social Responsibility

- Advance social responsibility education in interdisciplinary fields that contribute to information integrity research.
- Incorporate ethical, legal, and social implications panels and community advisory boards into research programs, especially for programs that affect information consumption, safeguards, or platform protections.
- Increase support for computer scientists studying information integrity to integrate more fully social science models of an individual's behavior and beliefs.
- Develop and expand metrics and algorithmic accountability approaches to assist in the design of socially responsible information technologies.
- Raise awareness of best practices for multidisciplinary research on social responsibility and mechanisms to incorporate social responsibility objectives into use-inspired, translational, and outcomes research.

5.2.2 Expand and Accelerate Information Integrity Research

Progress toward information integrity research goals requires research funding sustained over the next several decades. The following are actions related to research programs and critical investments in infrastructure.

5.2.2.1 Research Programs

- Support programs and solicitations that align with R&D priorities, objectives, and actions outlined in Sections 3 and 4 of this Roadmap.
- Provide training for researchers and government agencies that conduct and fund research on information integrity in research design and causal inference to more rigorously evaluate the consequences of social media on U.S. citizens and global populations.
- Ensure that institutional review boards have adequate guidance to maintain protections for human subjects while also enabling high-impact research teams to operate at the speed of relevance.
- Create challenges and competitions around R&D activities in selected areas.
- Increase the visibility of research progress across governmental funding agencies.
- Support scientific organizations that are developing guidance on R&D priorities.

5.2.2.2 Support for Researchers and Partners

- Expand support resources and training programs to assist information integrity researchers that inadvertently become targets of manipulated-information campaigns or harassment.
- Support a community of practice across researchers and practitioners who share this unique occupational hazard.
- Create shared resources for organization leadership and communications teams who are likewise affected by manipulated-information campaigns to assist with response efforts, compassionate leadership of affected employees, and employee retention.

5.2.2.3 Research Infrastructure

- Support and extend research infrastructure that enables access to data, algorithms, computational resources, testbeds, and realistic experimental settings to conduct research.
- Explore partnerships with international digital media observatories and investigate how similar models might be established in the United States.
- Expand shared datasets and test beds to compare promising tools for understanding and addressing manipulated-information campaigns.
- Develop and refine shared measurement methodologies and instrumentation to enhance rigor across the field.

5.2.3 Increase the Effectiveness of Use-Inspired, Translational, and Outcomes Research

The field of information integrity will be most effective when it has strong connections with real-world applications, use cases, and evaluations. Partnerships will play a key role. Important actions are discussed in the following subsections.

5.2.3.1 Partnerships with Government Practitioners

- Work with government practitioners to develop representative use cases and datasets to drive fundamental and use-inspired research.
- Foster partnerships that embed researchers with government decision makers (e.g., information integrity analysts or policymakers) to evaluate and refine research prototypes in real-world settings.

- Create a community of practice for Federal practitioners who address information corruption to interact with researchers, form partnerships, share best practices, and inform researchers about emerging challenges.
- Explore mechanisms for technology accelerators and digital services to contribute to translation efforts.

5.2.3.2 Collaborations with Businesses and Entrepreneurs

- Coordinate and expand existing representative surveys of American public opinion to access the extent of people's support for providing researchers access to their social media data.
- Explore collaborations toward data transparency and algorithmic accountability to assist with translational and outcomes research.
- Encourage internships and sabbaticals in which information integrity researchers transfer research findings into practice.

5.2.3.3 Engagements with Communities

- Collect evidence to inform those working to mitigate the effects of corrupted information on best practices to engage with community members.
- Develop participatory research approaches that include community members in the design and execution of studies and provide them with easily digestible information about research outcomes they care about.

5.2.3.4 Engagements with International Partners

- Expand on existing efforts, exchange lessons learned, and further establish joint research and pilot programs with researchers and entities outside the United States.
- Expand partnerships that enable development of information integrity tools for low-resource languages and assessment of manipulated-information campaigns in parts of the world where research has traditionally been sparse.
- Assess establishing a partnership with the European Union (EU) to provide U.S. researchers with access to social media data accessible under the 2022 EU Code of Practice on Disinformation.[80]

6. Glossary

actor *(malicious)* An individual or a group posing a threat.[81]

channels Conduits for information flow, including television, print media, books, social media platforms, and word of mouth.

corrupted information An umbrella term for information content that is inaccurate, misleading, deceptive, or inauthentically amplified and, regardless of intent, risks harm to individuals, communities, institutions, or society at large.

counterspeech Opposing corrupted information with high-quality information.

debunking See *counterspeech*.

deep fake An image or video that has been convincingly altered and manipulated to misrepresent someone or present them as doing or saying something that they did not say or do.

digital literacy People's ability to find, evaluate, and communicate information on digital devices and platforms. May also refer to their ability to maintain their privacy and security online.

disinformation Corrupted information that contains false content and is created or spread intentionally with the purpose of altering a specific target audience's attitudes or behavior.

distorting amplification Inauthentic actions to manipulate the spread or reach of content to deceive audiences about its volume, balance, or public appeal. May involve techniques such as automated bots or troll farms.

doxing Publication of an individual's personally identifiable or sensitive information, or an organization's confidential information, and publishing it (typically online) for malicious purposes.

extended reality Realistic or real-appearing immersive virtual environment. Also "XR." Includes augmented-reality, mixed-reality, and virtual-reality technologies.

foreign malign influence A hostile effort undertaken by, at the direction of, or on behalf of or with the substantial support of, the government of a covered foreign country with the objective of influencing, through overt or covert means the political, military, economic, or other policies or activities of the United States Government or state or local governments, including any election within the United States, or public opinion within the United States.[82,83]

fundamental research Basic[84] and applied research[85] in science and engineering, the results of which ordinarily are published and shared broadly within the scientific community, as distinguished from proprietary research and from industrial development, design, production, and product utilization, the results of which ordinarily are restricted for proprietary or national security reasons.[86]

gray zone activities A set of activities that occur between peace (or cooperation) and war (or armed conflict). A multitude of activities fall into this murky in-between, from nefarious economic activities, influence operations, and cyberattacks to mercenary operations, assassinations, and disinformation campaigns.[87]

harm Negative impact, costs, or effects on individuals, groups, institutions, countries, and society, including damage to physical or mental health or safety, reputation and trust, financial and economic welfare, and risks to national security. May be indirect, such as false and partisan claims increasing polarization, which in turn may fracture community trust and cooperation. May also accrue to purveyors of misinformation as when they are exposed and lose public standing.

influence campaign A prolonged series of actions (e.g., messaging) to inform or influence a targeted population.

information advantage A state of relative advantage predicated on superiority of information collection and analysis that enables greater situational awareness, greater understanding of the operational environment, and correspondingly faster, more effective decision-making.

information ecosystem A dynamic and evolving sociotechnical system, comprising information producers and consumers, groups and social networks, technology, system dynamics, and social institutions and contexts, that provides content.

information integrity The quality or state of information and associated patterns of creation, exchange, and consumption in society along a spectrum from low integrity to high integrity, where high-integrity information is trustworthy; distinguishes fact from fiction, opinion, and inference; acknowledges uncertainties; and is transparent about its level of vetting.

information literacy The United Nations Educational, Scientific and Cultural Organization defines information literacy as the ability of "people in all walks of life to seek, evaluate, use and create information effectively to achieve their personal, social, occupational and educational goals." Health, scientific, digital, and media literacy are aspects of information literacy.[88]

inoculation Proactive, pre-exposure, or prebunking of corrupted information or an information campaign to build resistance to similar information or campaigns in the future, with the intention of preventing or mitigating its harmful effects on beliefs.

manipulated information A type of corrupted information that is deliberately distorted, inaccurate, misleading, deceptive, or deceptively amplified. May include altered text, audio, or video; distorted statistics; misleading visualizations of data; critical information omissions that distort content; and deceptive presentations of fact designed to mislead an audience.

malinformation A type of corrupted information that is accurate but is strategically spread or slanted to cause harm.

media literacy The capability to access, analyze, evaluate, create, and participate with messages in a variety of forms — from print to video to the internet. Media literacy builds an understanding of the role of media in society as well as essential skills of inquiry and self-expression necessary for citizens of a democracy.[89]

misinformation False, inaccurate, or incorrect information that departs from the facts or preponderance of evidence. May be innocuous. Falls under the umbrella of corrupted information only if it has the potential to cause harm.

narrative A spoken or written account of connected events; a story.

open information ecosystem An information ecosystem that allows for the free exchange of content, ideas, and connections.

outcomes research Systematic evaluation of the success or failure of a tool, strategy, proposed action, or system. May include evaluation of factors in success or failure.

platform *(usage in this report)* An application or website that serves as a base from which a digital service is provided.

polarization Divided and opposing clusters of belief, income, or other attributes of a population.

prebunking Preemptively debunking a line of disinformation by publishing an account of that disinformation along with a simultaneous refutation before the disinformation itself is actually disseminated by its author.[90]

recommender system An algorithm or set of algorithms that suggests or curates content based on data about past information consumption or content features. Employed by many social media platforms to determine what is highlighted in users' newsfeeds or to recommend what users should access or purchase.

resilience The ability to operate effectively in the face of adversity, to withstand and rapidly recover from disruptions, and to adapt to changing conditions.

socially responsible innovation Processes or artifacts developed in accordance with ethical, legal, and pro-social principles.

swatting A false report of an ongoing emergency or threat of violence intended to prompt an immediate tactical law enforcement response, typically by a SWAT (Special Weapons and Tactics) team, to a specific location. May be motivated by revenge, used as a form of harassment, or used as a prank, but it is a crime that may have potentially deadly consequences.

translational research Research whose goal is to produce more meaningful, applicable results that directly benefit human health or other societal goals.

trolling (Internet slang) insincere, digressive, inflammatory, aggravating, or malicious online behavior characterized by aggressive and deliberate provocation of others.

use-inspired research Research whose rationale, conceptualization, and directions are driven by the potential use to which the knowledge will be put.

Appendix A. Bipartisan Executive and Congressional Actions

Addressing the challenges associated with information integrity has yielded bipartisan efforts over multiple administrations. These bipartisan efforts focus on mitigating threats from foreign adversaries and harms to the American public, and these policy goals help to frame the harm-mitigation emphasis within the Roadmap.

Threats from foreign propaganda and disinformation. Countering disinformation campaigns by America's adversaries has led to notable bipartisan efforts to:

- **Counter foreign disinformation that aims to undermine U.S. national security interests:** The U.S. Department of State's Global Engagement Center was established by the Obama Administration[91] and the bipartisan work of Senator Portman and Senator Murphy[92] to lead and coordinate Federal Government efforts that counter foreign disinformation aimed at undermining U.S. national security interests.
- **Counter foreign malign influence and covert distribution of disinformation that aims to undermine elections:** The National Cyber Strategy of the USA[93] and Executive Order 13848: Imposing Certain Sanctions in the Event of Foreign Interference in a United States Election,[94] released by the Trump Administration and extended by the Biden-Harris Administration, establish a priority action to counter online malign influence and disinformation, and declare that interference in United States elections by covert disinformation constitutes an "extraordinary threat to the national security."

Protection against scams. Consumer harms perpetrated through disinformation spread online has been the source of a continued congressional hearings.

- **Addressing financial fraud that targets veterans:** In 2019, the House Committee on Veterans' Affairs held a bipartisan hearing that examined malicious actors who impersonated veterans' groups to target veterans with benefits scams and propaganda[95] and probed mechanisms for fraud that resulted in a reported $66 million in losses among military retirees and veterans.[96]
- **Combating scams related to COVID-19:** The Senate Commerce Committee Subcommittee on Consumer Protection, Product Safety, and Data Security held a bipartisan hearing to raise awareness of ongoing scams related to COVID-19, such as identity theft, fake personal protective equipment, and fraudulent vaccination claims. [97] In opening remarks, Sen. Blumenthal (D-CT) emphasized the bipartisan nature of the hearing and the committee's efforts to protect consumers.[98]

Harms to children and youth. A multitude of congressional hearings have focused on the impacts to teens from social media platform recommendation algorithms such as:

- **Eating disorders:** A bipartisan group of senators expressed their concerns regarding social media's impact in promoting eating disorders among its younger users.[99]
- **Harmful behavior:** A series of hearings by the Senate Commerce Subcommittee on Consumer Protection, Product Safety, and Data Security focused on the misuse of social media applications to harm kids, promote destructive acts and deadly challenges, bullying, manipulative influencer marketing, and grooming.[100]

Elder fraud. In 2022, Congress enacted several bipartisan efforts to address elder fraud:

- **Preventing scams:** The Fraud and Scam Reduction Act[101] established a Scams Against Older Adults Advisory Group to prevent scams that target seniors and an office within the Bureau of Consumer Protection to monitor mail, television, internet, and telemarketing for fraud targeting seniors.
- **Empowering communities to fight fraud:** A hearing by the U.S. Senate Special Committee on Aging examined harms to seniors from fraudulent information and mechanisms to empower communities.[102] Both Ranking Member Sen. Tim Scott (R-SC) and Chairman Sen. Bob Casey (D-PA) emphasized bipartisan commitment to assisting aging Americans. [103,104]

Appendix B. Information Ecosystems Characterization

High- and low-integrity information exists within complex, multi-layered, global, evolving networks of information ecosystems, composed of many participants with different roles, motivations, and technology preferences and capabilities. Information ecosystems have developed over centuries of human existence, from early channels of person-to-person communication along trade routes, to the invention and content of books, to today's highly digital ecosystems. Today's ecosystems include many technological media tools (developed by organizations with their own incentives) that catalyze and shape the creation, flow, and consumption of information in society. Information ecosystems include the following elements and may include or come to include others.

Channels. Information across the integrity spectrum may flow through diverse channels, including television, print media, books, social media platforms, and word of mouth. Lesser studied conduits for corrupted information include book reviews, cartoons, multiplayer online games, or paid advertisements that appear to be legitimate news. The channels for manipulated information continue to evolve rapidly with the advent of new innovations in platform technologies, and augmented and virtual reality.

Contexts. Real-world events affect incentives around information manipulation as well as the scope, speed, and significance of its effects. The nature of manipulation varies across naturally occurring events such as disasters, geopolitical events such as the war in Ukraine, and biological events such as disease outbreaks. For example, the global pandemic increased economic incentives and motivated technological mechanisms to make online shopping easy, which reduced the difficulties of obtaining goods and services but also opened the door wider to the sale of fraudulent medications. Context likewise affects the impact of corrupted information, where the same idea can be satire in one context or become deadly advice within another.

Entities. Information ecosystems are populated with persons, groups (families, neighborhoods, social media groups), organizations (including corporations, online enterprises, national agencies, and international collaborations), and societal institutions (including those with community, educational, health, geopolitical, and commercial missions).

Interactions. Wide-ranging social, professional, commercial, familial, cultural, and geographical interactions affect information exposure in unpredictable ways. Interactions occur across social networks, interests, languages, cultures, and borders. Engagement with technical components creates intended and unintended effects in the creation, prevalence, and persistence of corrupted information.

Motivations and incentives. Entities may have a number of motives to manipulate information. They may be motivated by the potential of financial profit (as in the sale of fraudulent goods and services), or they may have reputational motives (to develop a reputation for a skill), geopolitical motives (to persuade citizens a war is justified), or motives for group cohesion (to recruit or retain group members). Entities also may spread false information accidentally, without ascertaining whether the information is true or false, sometimes because they have poor access to correct sources, or because they do not care. Entities may also use false information to justify hatred against specific groups, fuel ideological allegiance to their own perceived group, or rationalize harm to other persons or groups.[105]

Roles. Information manipulation is often framed in terms of competitors and adversaries who manipulate information and subjects who are the recipients of manipulated information. However, information ecosystems are fluid, and entities may fill different roles at different times. Depending on context, entities may be influenced by manipulation, may make conjectures that become corrupted information, may spread corrupted information (sometimes unintentionally), or may attempt to counter corrupted information. Similarly, technology platforms may sometimes provide transparency about the popularity of an idea, or become the target of manipulation.

Technical Components. Technology modulates the creation, adaptation, visibility, and dissemination of both high- and low-integrity information. Artificial intelligence components (such as recommender systems, amplification algorithms, microtargeting, and content moderation algorithms) affect what information spreads, to whom, and how quickly. These components influence the search terms used to acquire information and the relative visibility of both content and its apparent public support. In addition, manipulation technologies are continually advancing, enabling the rapid generation of multimodal falsifications, so-called "bots" that communicate through language models, and automation to distort distribution.

Appendix C. Harms That Warrant Further Research

Mitigating harms and addressing factors within information ecosystems that increase risks to people and populations are among the key objectives of this Roadmap, and context about the potential for harm is relevant when formulating research questions. This appendix provides additional detail and references about the harms introduced in Section 2.2.

C.1 Harms to Consumers and Companies

Financial harm to consumers. Financial harm from deceptive ads and other manipulated information on social media platforms is a growing problem, creating substantial consumer injury. In 2021, more than 95,000 people reported about $770 million in losses due to fraud initiated on social media platforms (18 times more than what was reported in 2017).[106] In addition, corrupted medical information can place a financial burden on consumers when it leads to degradation of health outcomes, loss of wages, or higher treatment costs.

Harms to corporations. State and non-state actors are using corrupted information to wage economic warfare or manipulate consumer confidence in markets. Large corporations increasingly find that traditional brand management strategies are insufficient to address the risks posed by information manipulation.[107]

Service disruptions. False claims can lead to destruction of corporate assets, damage to critical infrastructure, and service disruptions. For example, during 2020, false claims that 5G internet caused the coronavirus spread online, causing people to set fire to over 70 cell phone towers in the United Kingdom and attack the engineers who fix and maintain the equipment.[108]

C.2 Harms to Individuals and Families

Elder fraud. The number of older Americans falling victim to internet crimes like romance scams, investment fraud, healthcare miracle cures, government impersonation, and tech support fraud, and the cost of the associated losses, has risen dramatically with the rise of broadband connectivity.[109] Disinformation and other forms of information manipulation, to include misrepresentation of identity, are central to this criminal activity that disproportionally impacts senior citizens.[110] According to the National Elder Fraud Hotline, losses were over $184 million in 2018 alone.[111]

Harms to youth. Children and youth constitute a population especially vulnerable to corrupted information. Recognizing that it is not always within the cognitive and emotional faculties of a child to discern between reliable and unreliable information, study conclusions invite acute concern about the harmful effects of false and misleading information on children. Up to three-quarters of children reported feeling unable to judge the veracity of the information they encounter online—a finding especially true among young children.[112] U.S. teens aged 13–17 get their news more frequently from social media sites than directly from news organizations, and 60 percent of teens who get news from YouTube say they are more likely to get it from celebrities, influencers, and personalities than from news organizations (39 percent).[113] The UK's Commission on Fake News and Critical Literacy in Schools concluded that only two percent of children and young people have the critical literacy skills they need to judge whether a news story is real or false.[114] The immediate consequences of this susceptibility, measured in terms of jeopardized psychological well-being and distrust of authoritative information, are complemented by more enduring perils to generational faith in democracy.

Health. Corrupted health information can contribute to physical and mental distress, injury, disability, or death, especially when it causes individuals to refuse life-saving treatments or to take actions detrimental to their own health. A stark example of the harm of corrupted health information is the 2021 public advisory by the U.S. Surgeon General warning the public about the urgent threat of health misinformation, especially for COVID-19.[115] Patterns persist across disease types, with false beliefs that HIV does not cause AIDS leading to the death of at least 300,000 people in Africa[116] and content falsely describing eating disorders as a "lifestyle choice" contributing to teen anorexia.[117] Dangerous suggestions, such as the "blackout challenge," create yet another vehicle for injury or death.[118] As corrupted health information affects larger populations, costs to public health, disability services, and Medicare likewise become a major public health concern with significant financial and societal costs.

Personal agency. By design, information manipulation seeks to alter a person's beliefs and actions. Harm is often created when those beliefs or actions are not what the person would have chosen in the absence of the manipulation. Beyond the direct harms that can result from manipulation (such as exploitation, unfairness, and the deprivation of benefits), the deeper harm is infringement of individual autonomy, which can have far-reaching negative effects on democratic societies.[119] Examples of threats to autonomy include corrupted information campaigns that impede free speech and that discourage civic engagement, protesting injustice, or seeking professional mental health help.

Safety and reputation. Broad safety concerns including harassment, child exploitation, cyberbullying, and violent extremism, as well as reputation concerns, defamation, and invasion of privacy, are among the key issues motivating research in information manipulation. In its 2021 national survey, the Anti-Defamation League's Center for Technology and Society reports that 41 percent of Americans say they experienced harassment on social media, and 27 percent of respondents say they experienced severe and sustained harassment comprising sexual harassment, stalking, physical threats, swatting, or doxing.[120] The report by Plan International, which surveyed over 14,000 girls and young women in 31 countries, indicates that more than half of respondents had been harassed online, and that one in four felt physically unsafe as a result, indicative of the threats to safety and reputation that online harassment poses to women and girls around the globe.[121] Online harassment often involves the use of false or manipulated information, and it disproportionally affects people based on their gender, race or ethnicity, or religion. This includes image-based abuse and the non-consensual distribution of intimate images, including malicious deepfakes, which are disproportionately generated in the form of non-consensual, nude images of women. Furthermore, the Anti-Defamation League also reports that political views are perceived as the top reason for harassment, raising questions about the relationship between corrupted information and politically-motivated harassment, especially as it relates to public servants, healthcare workers, educators, and others who work in service positions.

C.3 Harms to National Security

National preparedness. Corrupted information may have significantly destructive impacts on the state of U.S. preparedness to sustain national critical infrastructures and respond to rapidly evolving events such as disasters and emergencies. The systems that underwrite the national infrastructures that support communications, critical manufacturing, emergency services, food and agriculture, financial services, healthcare, transportation systems, and water and waste systems are susceptible to corrupted information that can produce urban road network disruptions, spread rumors about the safety of air travel, agitate volatility in the stock market, and increase commodity consumption (hoarding) behavior.[122,123,124,125] Additionally, the spread of corrupted information can have a variety of impacts within the emergency management and homeland security environment, to include operational disruptions, diminished public trust and confidence, and offline incidents that can threaten responder and civilian safety and security during emergency situations.

Rule of law. Corrupted information compounds such already onerous rule-of-law challenges as human trafficking, illegal immigration, and terrorism by straining limited resources and inducing the participation and involvement of individuals who might not otherwise be inclined to engage in criminal activity. The resources that key non-governmental organizations and law enforcement bodies rely on to counter human trafficking and mitigate its impact are strained when untrue narratives inundate them with false reports.[126, 127, 128] False information that draws illegal border crossers to the United States aggravates the frequently dangerous societal and legal impacts of illegal immigration.[129] Corrupted information has been used by international terrorist organizations to garner sympathy, material support, and direct participation in terrorist activities. The comparable use of corrupted information by organized criminal groups and violent extremists grew exponentially during the COVID-19 pandemic.[130, 131, 132]

Strategic miscalculation. Information advantage is fundamental to military preparedness and sound decision-making. National security policies stemming from corrupted information, be it a function of purposely falsified intelligence or the outraged demands for action from a deliberately misled public, can be disastrous to a nation's strategic interests. Corrupted information may be leveraged by a foreign adversary to compel military action from within, damaging the targeted country and its allies or conferring advantage to the adversary.[133]

Widespread civil unrest. Corrupted information has been employed globally as a means of attempting to destabilize and fracture societies along political, religious, and racial fissures.[134] The proliferation of inexpensive and easily accessed communication mediums empowers state and non-state actors alike with the means to inject confusion, grievance, and rage directly into the social fabric of a community, city, or country, on a scale and with a rapidity that eclipses that of any preceding methodology. The result is an acutely heightened risk that ideologically opposed groups are susceptible to being manipulated into violent and potentially sustained confrontations. One important example is the wide-ranging death and destruction of mosques led by the Sinhalese Buddhist-majority in Sri Lanka in early 2018, actions that have been linked to social media activity. In a publicized response to a 2020 report detailing this civil unrest, Facebook acknowledged that content shared via Facebook and WhatsApp "may have helped stoke the violence," and in a subsequent statement apologized for "the very real human rights impacts that resulted."[135,136]

C.4 Harms to Society and Democratic Processes

Correlated escalation from online content to offline violence. A growing body of research explores the relationships between online polarizing content, the nature of online communication, and offline violence. [137, 138] Examples of correlated violence include the 2018 Burma genocide resulting from an information campaign orchestrated by Burma's military[139] and racially motivated anti-refugee attacks across Germany linked more strongly to Facebook usage than political ideology.[140] Nevertheless, additional research is needed to understand causal connections from online corrupted information to offline violence, particularly as it relates to politically-motivated harassment, attacks on women, and attacks on other marginalized groups.

Distrust in elections. Because democratic satisfaction and participation are closely associated with the perceived legitimacy of election outcomes, the actual and rumored information manipulation over elections poses a significant threat to democracy.[141]

Distrust of the media and threats to free press. The ability to both falsify information and present it in formats that look or act like news can make it harder for people to distinguish between sources and articles that follow journalistic standards and those that do not. Changes to the nature of journalism and consolidation of media outlets have further undermined trust. Attacks against journalists have become more prevalent, and many countries have enacted "fake news"[142] or "counter-disinformation" legislation[143] to suppress freedom of the press. Overall, government censorship of the media remains a leading indicator for countries in a state of democratic decline.[144]

Oppression of women and other marginalized groups. Manipulated information on social media often involves targeted harassment campaigns that seek to silence and marginalize specific groups in society, and to make it appear that actors spreading manipulated information are conveying an outsized consensus. Targeted use of trolling,[145] doxing, and other techniques has intentionally resulted in silencing voices and suppressing civic participation. For example, a growing body of research indicates that, globally, women in politics, the media, or even the U.S. military have been disproportionately targeted by manipulated-information campaigns that frame women as weak, untrustworthy, or incapable of holding office.[146]

Polarization. Information manipulation is used to reinforce insular and mutually suspicious online communities in which false claims are repeated, magnified, and considered as facts, making discourse between different groups of people more difficult. Once established, polarization provides fertile ground for social manipulation.

Endnotes

[1] Department of Justice (n.d.). National elder fraud hotline. *Office for Victims of Crime*. DOJ. https://ovc.ojp.gov/program/stop-elder-fraud/providing-help-restoring-hope

[2] Federal Communications Commission (2021). Veterans targeted in benefits scams. *Consumer*. FCC. https://www.fcc.gov/veterans-targeted-benefits-scams

[3] lesbian, gay, bisexual, transgender, queer, and intersex

[4] Sumpter, D. (2018). *Outnumbered: From Facebook and Google to fake news and filter-bubbles - the algorithms that control our lives*. Bloomsbury Publishing.

[5] Knutson, R. (2021). The outrage algorithm (No. 4). In *The Facebook files*. *Wall Street Journal*. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f

[6] Jackson, M. O., et al. (2022). Learning through the grapevine and the impact of the breadth and depth of social networks. *PNAS, 119*(34). https://doi.org/10.1073/pnas.2205549119

[7] Bail, C. (2021). *Breaking the Social Media Prism: How to make our platforms less polarizing*. Princeton University Press. https://doi.org/10.2307/j.ctv18zhdhg

[8] The term *service agency* is used to denote Federal agencies and organizations that have the authority to take actions or create guidance that affects policy or citizens, either nationally or internationally. It is defined in contrast to research agencies. Service agencies include regulatory agencies, law enforcement agencies, military services, the Department of State, consumer protection agencies, and public health agencies.

[9] The White House (2021). Fact sheet: The Biden-Harris Administration is taking action to restore and strengthen American democracy. *Briefing Room*. White House. https://www.whitehouse.gov/briefing-room/statements-releases/2021/12/08/fact-sheet-the-biden-harris-administration-is-taking-action-to-restore-and-strengthen-american-democracy/

[10] See https://www.nitrd.gov/87-fr-15274-responses/ for details on the Request for Information.

[11] The White House (2016). DCPD-201600149 - Executive Order 13721—Developing an integrated global engagement center to support government-wide counterterrorism communications activities directed abroad and revoking Executive Order 13584. *GovInfo*. https://www.govinfo.gov/app/details/DCPD-201600149

[12] U.S. Congress (2016). Public Law 114-328, National Defense Authorization Act for fiscal year 2017. *Public Laws*. Congress. https://www.congress.gov/114/plaws/publ328/PLAW-114publ328.pdf

[13] The White House (2018). National cyber strategy of the United States of America. *Archives*. Trump White House. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf

[14] The White House (2018). Executive Order 13848—Imposing certain sanctions in the event of foreign interference in a United States election. *GovInfo*. https://www.govinfo.gov/content/pkg/DCPD-201800593/html/DCPD-201800593.htm

[15] Federal Communications Commission (2021). Veterans targeted in benefits scams. *Consumer*. FCC. https://www.fcc.gov/veterans-targeted-benefits-scams

[16] U.S. Senate (2021). Curbing COVID Cons: Warning consumers about pandemic frauds, scams, and swindles. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/4/curbing-covid-cons-warning-consumers-about-pandemic-frauds-scams-and-swindles

[17] U.S. Senate (2021). Curbing COVID Cons: Warning Consumers about Pandemic Frauds, Scams, and Swindles. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/4/curbing-covid-cons-warning-consumers-about-pandemic-frauds-scams-and-swindles

[18] Shelley Moore Capito (2021). Senators concerned eating disorders are promoted on social media, pen letter to Facebook. *News*. Capito. https://www.capito.senate.gov/news/in-the-news/senators-concerned-eating-disorders-are-promoted-on-social-media-pen-letter-to-facebook

[19] U.S. Senate (2021). Protecting Kids Online: Snapchat, TikTok, and YouTube. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/10/protecting-kids-online-snapchat-tiktok-and-youtube

[20] U.S. Congress (2022). Public Law 117-103, Consolidated Appropriations Act, 2022. *Public Laws*. Congress. https://www.congress.gov/117/plaws/publ103/PLAW-117publ103.pdf

[21] U.S. Senate (2022). Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/hearings/stopping-senior-scams-empowering-communities-to-fight-fraud

[22] Scott, T. (2022). Opening Statement: Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/imo/media/doc/Opening%20Statement_Scott%2009.22.2022.pdf.pdf

[23] Casey, B. (2022). Opening Statement: Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/imo/media/doc/Opening%20Statement_Casey%2009.22.2022.pdf

[24] The White House (2022). Fact sheet: United States and 60 global partners launch declaration for the future of the internet. *Briefing Room*. White House. https://www.whitehouse.gov/briefing-room/statements-releases/2022/04/28/fact-sheet-united-states-and-60-global-partners-launch-declaration-for-the-future-of-the-internet/

[25] The White House (2021). Fact sheet: The Biden-Harris Administration is taking action to restore and strengthen American democracy. *Briefing Room*. White House. https://www.whitehouse.gov/briefing-room/statements-releases/2021/12/08/fact-sheet-the-biden-harris-administration-is-taking-action-to-restore-and-strengthen-american-democracy/; The White House (2021). Announcing the Presidential Initiative for Democratic Renewal. *Briefing Room*. White House. https://www.whitehouse.gov/briefing-room/statements-releases/2021/12/09/fact-sheet-announcing-the-presidential-initiative-for-democratic-renewal/

[26] The White House (2021). National strategy for countering domestic terrorism. *Publications*. White House. https://www.whitehouse.gov/wp-content/uploads/2021/06/National-Strategy-for-Countering-Domestic-Terrorism.pdf

[27] The term *corrupted* has two definitions, one that involves dishonest actions for monetary or personal gain and another that involves unintentional alterations or errors. This term was selected to reflect the dual nature of information integrity challenges, because some manipulations occur intentionally for gain or profit while others occur accidentally when information becomes stale or is repeated in the wrong context.

[28] For example, the rise of "fake" civil societies or astroturfing: Page, M. T. (2021). Fake civil society: The rise of pro-government NGOs in Nigeria. *Publications*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2021/07/28/fake-civil-society-rise-of-pro-government-ngos-in-nigeria-pub-85041

[29] See, for example, Cybersecurity & Infrastructure Security Agency. Mis, dis, malinformation. *Election Security*. CISA. https://www.cisa.gov/mdm

[30] Department of Justice. National elder fraud hotline. *Office for Victims of Crime*. DOJ. https://ovc.ojp.gov/program/stop-elder-fraud/providing-help-restoring-hope

[31] U. S. Senate Select Committee on Intelligence (2019). Report on Russian active measures campaigns and interference in the 2016 U.S. election, vol. 2, p. 1. *Publications*. U.S. Senate. https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures

[32] U.S. Department of State (2022). Secretary Antony J. Blinken on the genocide and crimes against humanity in Burma. *Press Releases*. U.S. Department of State. https://www.state.gov/secretary-antony-j-blinken-at-the-united-states-holocaust-memorial-museum/

[33] Guzman, J. (2020). More than 70 cell phone towers in the UK have been set on fire due to 5G coronavirus conspiracy theory. *Changing America*. *The Hill*. https://thehill.com/changing-america/well-being/longevity/496502-more-than-70-cell-phone-towers-in-the-uk-have-been-set/

[34] Taub, A., et. al. (2018). Facebook fueled anti-refugee attacks in Germany, new research suggests. *The New York Times*. https://www.nytimes.com/2018/08/21/world/europe/facebook-refugee-attacks-germany.html

[35] Epstein, R. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcome of elections, *PNAS, 112*(33). https://www.pnas.org/doi/10.1073/pnas.1419828112

[36] Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media*. Princeton University Press. https://doi.org/10.2307/j.ctv8xnhtd

[37] See, for example, Gordon, J. (1997). John Stuart Mill and the 'Marketplace of Ideas.' *Social Theory and Practice*, *23*(2), 235–249. https://www.jstor.org/stable/23559183

[38] Resilience (n.d.). *Glossary*. CSRC. https://csrc.nist.gov/glossary/term/resilience

[39] The White House (2021). Memorandum on restoring trust in government through scientific integrity and evidence-based policymaking. *Briefing Room*. White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/memorandum-on-restoring-trust-in-government-through-scientific-integrity-and-evidence-based-policymaking/; The White House (2022). Fact sheet: Biden-Harris Administration launches year of evidence for action to fortify and expand evidence-based policymaking. *News*. White House. https://www.whitehouse.gov/ostp/news-updates/2022/04/07/fact-sheet-biden-harris-administration-launches-year-of-evidence-for-action-to-fortify-and-expand-evidence-based-policymaking/

[40] Such theories and frameworks include the ABCDE framework, which includes Actors, Behaviors, Content, Degree, and Effects; the Disinformation Analysis & Risk Management (DISARM) framework; Diffusion of Information Theory; the SCOTCH framework, which includes Sources, Channels, Objectives, Targets, Compositions, and Hooks of deliberate influence operations; the NATO framework, which describes the severity of risk based on Reach and Thematic Issues and frameworks of the Council of Europe, International Telecommunication Union (ITU), and the United Nations Educational, Scientific and Cultural Organization (UNESCO).

[41] Zmigrod, L., et al. (2021). The cognitive and perceptual correlates of ideological attitudes: A data-driven approach. Philosophical Transactions of the Royal Society B. The Royal Society Publishing. https://doi.org/10.1098/rstb.2020.0424

[42] Montrey, M., & Shultz, T. R. (2022). Copy the in-group: Group membership trumps perceived reliability, warmth, and competence in a social-learning task. *Psychological Science, 33*(1). https://doi.org/10.1177/09567976211032224

[43] Mosleh, M., et al. (2021). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *PNAS, 118*(7). https://doi.org/10.1073/pnas.2022761118

[44] Hornsey, M. J., et al. (2020). Why facts are not enough: Understanding and managing the motivated rejection of science. *Current Directions in Psychological Science, 29*(6). https://doi.org/10.1177/0963721420969364

[45] Barrett, P. M., et al. (2021). Fueling the fire: How social media intensifies U.S. political polarization - and what can be done about it. *Static Pages*. Squarespace. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/613a4d4cc86b9d3810eb35aa/1631210832122/NYU+CBHR+Fueling+The+Fire_FINAL+ONLINE+REVISED+Sep7.pdf

[46] V-Dem Institute (2022). *Democracy Report 2022: Autocratization Changing Nature?* V-Dem Institute. https://v-dem.net/media/publications/dr_2022.pdf

[47] McPherson, M., et al. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

[48] Kennedy, E. H., et al. (2021). Morality, emotions, and the ideal environmentalist: Toward a conceptual framework for understanding political polarization. *American Behavioral Scientist, 66*(9). https://doi.org/10.1177/00027642211056258; Serrano-Puche, J. (2021). Digital disinformation and emotions: exploring the social risks of affective polarization. *International Review of Sociology, 31*(2), 231–245. https://www.tandfonline.com/doi/abs/10.1080/03906701.2021.1947953; McLaughlin, B., et al. (2019). Emotions and affective polarization: How enthusiasm and anxiety about presidential candidates affect interparty attitudes. *American Politics Research, 48*(2). https://doi.org/10.1177/1532673X19891423

[49] Levchenko, K., et al. (2011). Click trajectories: End-to-end analysis of the spam value chain. *IEEE Symposium on Security and Privacy*, 431–446. https://cseweb.ucsd.edu//~savage/papers/Oakland11.pdf

[50] Mooijman, M., et al. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour, 2*(6), 389–396. https://doi.org/10.1038/s41562-018-0353-0

[51] Southwell, B. G., et al. (2022). Defining and measuring scientific misinformation. *The ANNALS of the American Academy of Political and Social Science, 700*(1), 98–111. https://doi.org/10.1177/00027162221084709

[52] Vazire, S., et al. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162–168. https://doi.org/10.1177/09637214211067779

[53] Bilali, R. (2022). Fighting violent extremism with narrative intervention: Evidence from a field experiment in West Africa. *Psychological Science, 33*(2), 184–195. https://doi.org/10.1177/09567976211031895

[54] Traberg, C. S., et al. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science, 700*(1), 136–151, https://doi.org/10.1177/00027162221087936

[55] Wasserman, H., & Madrid-Morales, D. (Eds.) (2021). *Disinformation in the global south*. Wiley.

[56] Sumpter. D. (2018). *Outnumbered: From Facebook and Google to fake news and filter-bubbles – the algorithms that control our lives*. Bloomsbury Publishing.

[57] Salganik et al. Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5762). https://www.science.org/doi/full/10.1126/science.1121066

[58] Knutson, R. (2021). The outrage algorithm (No. 4). In *The Facebook files*. *Wall Street Journal*. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f; Kantrowitz, A. (2021). The case to reform the share button, according to Facebook's own research. *Big Technology*. https://bigtechnology.substack.com/p/the-case-to-reform-the-share-button?s=r

[59] Knutson, R. (2021). The outrage algorithm (No. 4). In *The Facebook files*. *Wall Street Journal*. https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f; Kantrowitz, A. (2021). The case to reform the share button, according to Facebook's own research. *Big Technology*. https://bigtechnology.substack.com/p/the-case-to-reform-the-share-button?s=r

[60] Jackson, M. O., et al. (2022). Learning through the grapevine and the impact of the breadth and depth of social networks. *PNAS, 119*(34). https://doi.org/10.1073/pnas.2205549119

[61] Hagey, K., & Horwitz, J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *Articles*. *Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline

[62] Bail, C. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

[63] Latapí Agudelo, M. A., et al. (2019). *International Journal of Corporate Social Responsibility, 4*(1). https://doi.org/10.1186/s40991-018-0039-y

[64] Hess, D. J., et al. (2021). A comparative, sociotechnical design perspective on responsible innovation: Multidisciplinary research and education on digitized energy and automated vehicles. *Journal of Responsible Innovation, 8*(3), 421–444. https://www.tandfonline.com/doi/full/10.1080/23299460.2021.1975377

[65] Barrett, P. M., et al. (2021). Fueling the fire: How social media intensifies U.S. political polarization - and what can be done about it. *Static Pages*. Squarespace.

https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/613a4d4cc86b9d3810eb35aa/1631210832122/NYU+CBHR+Fueling+The+Fire_FINAL+ONLINE+REVISED+Sep7.pdf

[66] Preece, J., & Maloney-Krichmar, D. (2017). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication, 10*(4). https://doi.org/10.1111/j.1083-6101.2005.tb00264.x

[67] Fisher, M. (2021). Disinformation for hire, a shadow industry, is quietly booming. *New York Times*. https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html

[68] Wasserman, H., & Madrid-Morales, D. (Eds.) (2021). *Disinformation in the Global South*. Wiley.

[69] Harmful information campaigns occurs when a person, group of people, or entity (a "threat actor") coordinate to distribute corrupt (false or misleading) information while concealing the true objectives of the campaign. The objectives of such campaigns can be broad (e.g., sowing discord in a population) or targeted (e.g., propagating a counternarrative to domestic protests) and may employ all information types (disinformation, misinformation, malinformation, propaganda, and true information). The target of a harmful information campaign is the person or group the threat actor aims to influence in order to achieve the campaign's objective. See https://www.dhs.gov/sites/default/files/publications/ia/ia_combatting-targeted-disinformation-campaigns.pdf

[70] Franconeri, S. L., et al. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest, 22*(3). https://doi.org/10.1177/15291006211051956; Reyna, V. F. (2020). A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *PNAS, 118*(15). https://doi.org/10.1073/pnas.1912441117

[71] Trust & Safety Team (n.d.). Keep your data safe. *About the Bureau*. U.S. Census Bureau. https://www.census.gov/about/trust-and-safety.html

[72] For example, the statistical data protections within The Foundations for Evidenced-based Policymaking Act of 2018.

[73] Such as the HL7 FHIR Accelerators, in which members of the private sector, research community, nonprofits, and government collaborate toward technology prototypes and standards to ease data access.

[74] Coalition for Content Provenance and Authenticity (n.d.). Overview. *C2PA*. https://c2pa.org/

[75] Partnership for Countering Influence Operations. Carnegie Endowment for International Peace. https://carnegieendowment.org/specialprojects/counteringinfluenceoperations

[76] European Digital Media Observatory. https://edmo.eu/

[77] European Commission (n.d.). A growing threat to European democracies. A strengthened EU code of practice on disinformation. *Priorities*. European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan/strengthened-eu-code-practice-disinformation_en

[78] Bureau of Economic Analysis (n.d.). Special sworn researcher program. *Research*. BEA. https://www.bea.gov/research/special-sworn-researcher-program

[79] For example, very simple algorithms—such as inverse popularity ranking—may affect short- and long-term choices in different ways, causing perceptions of popularity to decouple choices from general appeal in the short term (see, for example, Salganik, M. J., et al. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly, 71*(4). https://journals.sagepub.com/doi/abs/10.1177/019027250807100404).

[80] European Commission (n.d.). A growing threat to European democracies. A strengthened EU code of practice on disinformation. *Priorities*. European Commission. https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan/strengthened-eu-code-practice-disinformation_en

[81] Threat actor (n.d.). *Glossary*. CSRC. https://csrc.nist.gov/glossary/term/threat_actor

[82] U.S. House of Representatives (2022). 50 USC 3059: Foreign malign influence response center. *U.S. Code*. U.S. House of Representatives. https://uscode.house.gov/view.xhtml?req=(title:50%20section:3059%20edition:prelim)

[83] The term *covered foreign country* means the following:

(A) The Russian Federation.

(B) The Islamic Republic of Iran.

(C) The Democratic People's Republic of Korea.

(D) The People's Republic of China.

(E) Any other foreign country that the Director of the Foreign Malign Influence Response Center determines appropriate.

[84] Experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts. May include activities with broad or general applications in mind, such as the study of how plant genomes change, but should exclude research directed towards a specific application or requirement, such as the optimization of the genome of a specific crop species; see White House (2018). Circular A-11: Preparation, submission, and execution of the budget. *OMB*. White House. https://www.whitehouse.gov/wp-content/uploads/2018/06/a11.pdf. Also see NSF's definition of R&D: R&D. Frascati Manual, 2.5. *Statistics*. NSF. https://www.nsf.gov/statistics/randdef/manual.htm.

[85] Original investigation undertaken in order to acquire new knowledge. Directed primarily toward a specific practical aim or objective; see The White House (2018). Circular A-11: Preparation, submission, and execution of the budget. *OMB*. White House. https://www.whitehouse.gov/wp-content/uploads/2018/06/a11.pdf.

[86] National Archives (1985). NSDD 189 National policy on transfer of scientific, technical and engineering information. *NSDDs*. NARA. https://catalog.archives.gov/id/6879779; Defense Advanced Research Projects Agency (n.d.). Fundamental research. *For Universities*. DARPA. https://www.darpa.mil/work-with-us/for-universities/fundamental-research

[87] Atlantic Council (2022, February 23). Today's wars are fought in the 'gray zone.' Here's everything you need to know about it. *Future of Tech Competition*. https://www.atlanticcouncil.org/blogs/new-atlanticist/todays-wars-are-fought-in-the-gray-zone-heres-everything-you-need-to-know-about-it/

[88] United Nations Educational, Scientific and Cultural Organization (2022). Information literacy. *Information for All Programme*. UNESCO. Information Literacy, https://www.unesco.org/en/ifap/information-literacy

[89] Media literacy: A definition and more (n.d.). *Center for Media Literacy*. MediaLit. https://www.medialit.org/media-literacy-definition-and-more

[90] Prebunking definition (n.d.). *CyberWire*. The CyberWire. https://thecyberwire.com/glossary/prebunking

[91] The White House (2016). DCPD-201600149 - Executive Order 13721—Developing an integrated global engagement center to support government-wide counterterrorism communications activities directed abroad and revoking Executive Order 13584. *GovInfo*. https://www.govinfo.gov/app/details/DCPD-201600149

[92] U.S. Congress (2016). Public Law 114-328, National Defense Authorization Act for fiscal year 2017. *Public Laws*. Congress. https://www.congress.gov/114/plaws/publ328/PLAW-114publ328.pdf

[93] The White House (2018). National cyber strategy of the United States of America. *Archives*. Trump White House. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf

[94] The White House (2018). Executive Order 13848—Imposing certain sanctions in the event of foreign interference in a United States election. *GovInfo*. https://www.govinfo.gov/content/pkg/DCPD-201800593/html/DCPD-201800593.htm

[95] U.S. House of Representatives (2019). Hijacking our heroes: Exploiting veterans through disinformation on social media. *Committee on Veterans' Affairs Hearing*. U.S. House of Representatives. https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=110183

[96] Federal Communications Commission (2021). Veterans targeted in benefits scams. *Consumer*. FCC. https://www.fcc.gov/veterans-targeted-benefits-scams

[97] U.S. Senate (2021). Curbing COVID Cons: Warning consumers about pandemic frauds, scams, and swindles. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/4/curbing-covid-cons-warning-consumers-about-pandemic-frauds-scams-and-swindles

[98] U.S. Senate (2021). Curbing COVID Cons: Warning Consumers about Pandemic Frauds, Scams, and Swindles. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/4/curbing-covid-cons-warning-consumers-about-pandemic-frauds-scams-and-swindles

[99] Shelley Moore Capito (2021). Senators concerned eating disorders are promoted on social media, pen letter to Facebook. *News*. Capito. https://www.capito.senate.gov/news/in-the-news/senators-concerned-eating-disorders-are-promoted-on-social-media-pen-letter-to-facebook

[100] U.S. Senate (2021). Protecting Kids Online: Snapchat, TikTok, and YouTube. *U.S. Senate Committee on Commerce, Science, & Transportation Subcommittee Hearing*. U.S. Senate. https://www.commerce.senate.gov/2021/10/protecting-kids-online-snapchat-tiktok-and-youtube

[101] U.S. Congress (2022). Public Law 117-103, Consolidated Appropriations Act, 2022. *Public Laws*. Congress. https://www.congress.gov/117/plaws/publ103/PLAW-117publ103.pdf

[102] U.S. Senate (2022). Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/hearings/stopping-senior-scams-empowering-communities-to-fight-fraud

[103] Scott, T. (2022). Opening Statement: Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/imo/media/doc/Opening%20Statement_Scott%2009.22.2022.pdf.pdf

[104] Casey, B. (2022). Opening Statement: Stopping senior scams: Empowering communities to fight fraud. *U.S. Senate Special Committee on Aging Hearing*. U.S. Senate. https://www.aging.senate.gov/imo/media/doc/Opening%20Statement_Casey%2009.22.2022.pdf

[105] Ullrich Ecker et al. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*. 1, 13–19, January 2022, https://doi.org/10.1038/s44159-021-00006-y

[106] Fair, L. (2022). Gold mine for scammers: Social media. *Business Blog*. Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2022/01/gold-mine-scammers-social-media

[107] Network Contagion Research Institute. The future of disinformation operations and the coming war on brands. *NCRI Reports*. NCRI. https://networkcontagion.us/wp-content/uploads/NCRI-%E2%80%93-The-Future-of-Disinformation.pdf

[108] Guzman, J. (2020). More than 70 cell phone towers in the UK have been set on fire due to 5G coronavirus conspiracy theory. *Changing America*. *The Hill*. https://thehill.com/changing-america/well-being/longevity/496502-more-than-70-cell-phone-towers-in-the-uk-have-been-set/

[109] Federal Bureau of Investigation (2022). Elder fraud report 2021. *Internet Crime Complaint Center*. FBI. https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3ElderFraudReport.pdf

[110] Markowitz, A. (2022). Older Americans' cybercrime losses soared to $3 billion in 2021. *Money*. AARP. https://www.aarp.org/money/scams-fraud/info-2022/fbi-elder-fraud-report.html

[111] Department of Justice. National elder fraud hotline. *Office for Victims of Crime*. DOJ. https://ovc.ojp.gov/program/stop-elder-fraud/providing-help-restoring-hope

[112] Livingstone, S., et al. (2019). Global kids online comparative report. *Innocenti Research Report*. UNICEF. https://www.unicef-irc.org/publications/1059-global-kids-online-comparative-report.html

[113] Common Sense Media (2019). New survey reveals teens get their news from social media and YouTube. *Press Releases*. Common Sense. https://www.commonsensemedia.org/press-releases/new-survey-reveals-teens-get-their-news-from-social-media-and-youtube

[114] National Literacy Trust (2018). Fake news and critical literacy: The final report of the Commission on Fake News and the teaching of critical literacy in schools. *Media*. Literacy Trust UK. https://cdn.literacytrust.org.uk/media/documents/Fake_news_and_critical_literacy_-_final_report.pdf

[115] Department of Health and Human Services (2021). Confronting health misinformation. *Current Priorities of the U.S. Surgeon General*. HHS. https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf.

[116] Chigwedere, P., et al. (2008). Estimating the lost benefits of antiretroviral drug use in South Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes, 49*(4), 410–415, https://doi.org/10.1097/qai.0b013e31818a6cd5

[117] O'Sullivan, D. et al. (2021). Instagram promoted pages glorifying eating disorders to teen accounts. *Business*. CNN. https://www.cnn.com/2021/10/04/tech/instagram-facebook-eating-disorders/index.html

[118] People (2021). 10-year-old girl dies trying 'blackout challenge' from social media, mom says. *People*. https://people.com/human-interest/10-year-old-girl-dies-trying-blackout-challenge-from-tiktok/

[119] Susser, D. et al. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review, 4*(1). https://ssrn.com/abstract=3306006

[120] Anti-Defamation League (2022). Online hate and harassment: The American experience 2021. *Online Hate & Harassment*. ADL. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021

[121] Plan International (2020). Free to be online? Girls' and young women's experiences of online harassment. *State of the World's Girls*. Plan International. https://plan-international.org/uploads/2022/02/sotwgr2020-commsreport-en-2.pdf

[122] See https://www.cisa.gov/critical-infrastructure-sectors for a full list of 16 national critical infrastructure sectors.

[123] Waniek, M., et al. (2021). Traffic networks are vulnerable to disinformation attacks. *Scientific Reports, 11, 5329*. Nature. https://doi.org/10.1038/s41598-021-84291-w

[124] See Golovchenko, Y., & Adler-Nissen, R. (2018). Who spread disinformation about the MH17 crash? We followed the Twitter trail. *Monkey Cage*. *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2018/09/20/who-spread-information-disinformation-about-the-mh17-crash-we-followed-the-twitter-trail/ and Thomaselli, R. (2020). Man tries to delay flight by reporting fake bomb threat. *Airlines & Airports*. Travel Pulse. https://www.travelpulse.com/news/airlines/man-tries-to-delay-flight-by-reporting-fake-bomb-threat.html

[125] See Public Relations Society of America (2020). The financial drain of misinformation. *Though Leadership*. PRSA. https://prsay.prsa.org/2021/04/22/the-financial-drain-of-misinformation/

[126] U.S. Department of State (2022). 2022 trafficking in persons report. *Office to Monitor and Combat Trafficking in Persons*. U.S. Department of State. https://www.state.gov/wp-content/uploads/2022/10/20221020-2022-TIP-Report.pdf

[127] Rajan, A., et al. (2021). Countering QAnon: Understanding the role of human trafficking in the disinformation-extremist nexus. Polaris Project. https://polarisproject.org/wp-content/uploads/2021/02/Polaris-Report-Countering-QAnon.pdf

[128] Rajan, A. (2022). Separating fact from fiction can save trafficking victims and our democracy. *Opinion*. *The Hill*. https://thehill.com/opinion/national-security/589642-separating-fact-from-fiction-can-save-trafficking-victims-and-our/

[129] Guerrero, J. (2021). QAnon conspiracy theory: The border variant. *Opinion Column*. *Los Angeles Times*. https://www.latimes.com/opinion/story/2021-09-15/column-qanon-conspiracy-theory-the-border-variant

[130] United Nations Interregional Crime and Justice Research Institute (2020). Stop the virus of disinformation: The malicious use of social media by terrorist, violent extremist and criminal groups during the COVID-19 pandemic - November 2020. *Publications*. UNICRI. https://unicri.it/Publications/Malicious-use-cocialmedia-terrorists-extremists-criminals

[131] Brandon, S. (2021). Substance or snake oil? *Guides*. CREST. https://crestresearch.ac.uk/resources/substance-or-snake-oil/

[132] Innes, M. (2020). Soft facts and digital behavioural influencing after the 2017 terror attacks. *Reports*. CREST. https://crestresearch.ac.uk/resources/soft-facts-full-report/

[133] American Security Project (n.d.). Disinformation. *Public Diplomacy and Strategic Communication*. ASP. https://www.americansecurityproject.org/public-diplomacy-and-strategic-communication/disinformation/

[134] U. S. Senate Select Committee on Intelligence (2019). Report on Russian active measures campaigns and interference in the 2016 U.S. election, vol. 2, p. 1. *Publications*. U.S. Senate. https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures

[135] Facebook (2021). Sri Lanka human rights impact assessment. *Facebook Response*. Facebook. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf

[136] Yahoo Finance (2020). Facebook apologises for role in 2018 Sri Lanka unrest. *Finance*. Yahoo. https://finance.yahoo.com/news/facebook-apologises-role-2018-sri-lanka-unrest-081203917.html

[137] Gallacher, J. D., et al. (2021). Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters. *Social Media + Society, 7(1)*. https://doi.org/10.1177/2056305120984445

[138] Hassan, G., et al. (2018). Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International Journal of Developmental Science, 12*(1-2), 71–88. https://content.iospress.com/articles/international-journal-of-developmental-science/dev170233

[139] U.S. Department of State (2022). Secretary Antony J. Blinken on the genocide and crimes against humanity in Burma. *Press Releases*. U.S. Department of State. https://www.state.gov/secretary-antony-j-blinken-at-the-united-states-holocaust-memorial-museum/

[140] Taub, A., & Fisher, M. (2018). Facebook fueled anti-refugee attacks in Germany, new research suggests. *The New York Times*. https://www.nytimes.com/2018/08/21/world/europe/facebook-refugee-attacks-germany.html

[141] Epstein, R. & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS*, *112*(33). https://doi.org/10.1073/pnas.1419828112

[142] Wiseman, J. (2020). Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges. *International Press Institute*. https://ipi.media/rush-to-pass-fake-news-laws-during-covid-19-intensifying-global-media-freedom-challenges

[143] Yadav, K., et al. (2021). Countries have more than 100 laws on the books to combat misinformation. How well do they work? *The Bulletin of the Atomic Scientists*. https://thebulletin.org/premium/2021-05/countries-have-more-than-100-laws-on-the-books-to-combat-misinformation-how-well-do-they-work/

[144] V-Dem Institute (2022). *Democracy Report 2022: Autocratization Changing Nature?* https://v-dem.net/media/publications/dr_2022.pdf

[145] Trolling (internet slang): insincere, digressive, inflammatory, aggravating, or malicious online behavior, characterized by aggressive and deliberate provocation of others.

[146] Di Meco, L., & Wilfore, K. (2021). Gendered disinformation is a national security problem. *Brookings Institute*. https://www.brookings.edu/techstream/gendered-disinformation-is-a-national-security-problem/