Information Integrity
SaTC 2022 PI Meeting Breakout Group Report
Co-leads: Lisa Fazio, Benjamin Mako Hill, Kate Starbird
Scribe: Daniel Votipka

1. “Information Integrity” within Secure and Trustworthy Computing

Misinformation, disinformation, harassment, and strategic manipulation within online information spaces represent threats to a healthy society and functioning democracy. These challenges — which have emerged and continue to evolve at a complex intersection of technology, human psychology, culture, and politics — are an increasingly salient dimension of Secure and Trustworthy Computing.

The terms and definitions in this emerging field are still dynamic, with researchers, policy makers, and others still struggling to identify a common vocabulary and a shared understanding of what the problems are and what steps can be taken, by whom, to help mitigate these issues.

Within this dynamic space, SaTC should focus its attention around **mitigating** (identifying, hardening systems again, and reducing the impact of) **manipulation of information systems** (including content, networks, and algorithms) **which have harmful effects on individuals, groups, and society.**

2. The Study of Information Integrity should be Multi- and Interdisciplinary

Understanding these challenges will require a multi- and interdisciplinary approach. The NSF should support research from diverse perspectives and disciplines and incentivize collaborative, interdisciplinary teams to address the socio-technical problems. The portfolio should include work from the following disciplines (among others):

- **Information scientists** studying how false, misleading, or manipulative information spreads online;
- **Computer scientists** developing techniques for detection; designing solutions for algorithmic transparency
- **HCI researchers** looking at how the design of systems shapes how online tools are used to deceive and manipulate;
- **CSCW researchers**: looking at collaborative/participatory nature of online disinformation
- **Sociologists** looking at how the structure of systems (networks, algorithms) reflects and shapes influence operations
- **Education scholars** developing new digital and civic media literacies
- **Psychologists** studying why people are vulnerable to believing false information

- **Communication scholars** studying propaganda, historically and in modern, online systems

3. Key Research Challenges

Because this field is still emerging, our group identified many research challenges. We list a few key challenges here.

a. **Context matters:** Problems of Information Integrity (including mis- and disinformation as well as harassment) are highly *contextual*. This can make detection difficult and limit the efficacy of one-size-fits-all solutions.

b. **Take a broader lens:** Due to a number of factors, existing research on information integrity is limited to a small subsection of the problem — primarily English-language, primarily text, and primarily Twitter.

   i. **Support additional languages**: The problem of information integrity is a global one. Online harassment and disinformation are shaping political outcomes in Brazil, India, the Philippines, and elsewhere. Tactics used in one part of the world may be picked up and repurposed for use in another context. Studying what is happening in other parts of the world will be valuable for identifying tactics and determining how social and cultural factors shape these phenomena. Here in the United States, while English-language content is predominant, problems like misinformation, disinformation, and harassment also take place in other languages — where researchers are less likely to be looking and technology platforms are less likely to take action. The NSF should support research on information integrity that focuses on languages beyond English and locations beyond the United States.

   ii. **Support research on images, audio, and video:** Many of the most popular online platforms — e.g. Instagram, YouTube, and TikTok — focus on images and video. Even more traditionally text-based platforms such as Twitter incorporate images and video into their user experience. However, most research in the information integrity space focuses on textual content. The NSF should support research that enhances our understanding of how online misinformation, disinformation, manipulation, and harassment take shape within audio, video, and images. The NSF should also invest in the development of methods and research infrastructure for analyzing audio, video, and image content.

   iii. **Support cross-platform research:** Due to the public availability of content (and the text-based nature of the platform), a large portion of research into issues of information integrity online has focused on the Twitter platform. Though this research has provided important insights, it has considerable limitations. The field of information integrity needs additional research on other platforms, including popular platforms like YouTube and Facebook that have been understudied due to limited data access, and the long-tail of smaller platforms (some of which are especially relevant in this context). The field also needs more cross-platform research, as disinformation and harassment campaigns often use multiple platforms in complementary ways. The NSF should support research across diverse online platforms.

- **c. Support platform transparency initiatives:** Related to some of the concerns above, researchers, journalists, and policy makers have advocated for increased transparency by social media platforms — e.g. through intermediaries that manage access and safe harbor provisions. If possible, the NSF should also support efforts that increase platform transparency and provide broader access for researchers to social media data, while taking into account issues such as protecting user privacy.
- **d. Protect researchers.** A central concern for many researchers in this space is the safety and wellness of their research teams. Exposure to large volumes of social media content — especially conspiracy theories, disinformation, and abusive content — can have mental health impacts such as disorientation and depression. The NSF should invest in research to better understand these impacts and fund resources to support the mental health of researchers, especially student researchers, in this space. Additionally, for some topics such as disinformation and online harassment campaigns, the research space is an adversarial one, with the purveyors of those campaigns incentivized to undermine the research. Researchers in this space may be targeted for harassment, doxxing, and hack-and-leak operations. The NSF should consider organizing workshops and funding the development of resources to help researchers protect themselves.
- **e. Balance of basic and applied research.** Breakout group attendees argued that there continues to be a need for basic science to better understand emergent challenges around information integrity — even as there is a pressing need for solutions (e.g. platform designs, education, policy) to help mitigate these challenges. The NSF should continue to invest in a balanced portfolio that includes basic science and applied research in this space.

3. Promising Directions

The breakout group identified several promising directions in this space.

- a. **Understanding, documenting, and measuring harms.** Though there is increasing awareness around the proliferation of online misinformation, disinformation, harassment, and manipulation — and consequently, diminishing trust in information — we do not yet have a shared understanding of what the different harms might be. We recommend supporting emerging research that explores different kinds of harm — both direct and indirect, as well as at the individual, group, and societal levels — and develops frameworks for classifying and measuring those harms.
- **b. Closer collaboration with civil society groups.** The NSF should support collaborations between researchers and civil society organizations, for example within communities targeted and most harmed by misinformation, disinformation, manipulation, and harassment.
- c. **Support education and other individual-level interventions.** Researchers continue to debate the role of technology in facilitating and/or exacerbating the harms related to information integrity, and recent years have seen considerable debate about what social media platforms should or should not do to address these issues. Fewer resources have

been dedicated to educational interventions that could increase societal resilience to informational threats through, for example, improvements in information literacy. The NSF should support research exploring educational and individual-level interventions.

d. **Support research that embeds information literacy support into platform design.** One particular dimension that the NSF could support is design interventions, at the platform level, that support new media literacies — blending what we're learning about digital media literacy with platform design, giving users the signals they need to make better decisions about the information they consume and share.

e. **Learning from successful cases of online groups.** There are many cases of online communities successfully protecting themselves and handling information integrity issues well. The NSF should invest in research that studies how specific communities respond to informational threats, and help translate insights from those studies, for example about how certain moderation policies contribute to healthier online communities, into design insights and best practices.