Mary Ellen Zurko
Associate Editor in Chief

Disinformation and Reflections From Usable Security

By 2019, a number of security professionals in my sphere were questioning whether combatting misinformation was part of security or if it was entirely outside of our core area. Perhaps it was best treated as a potential application domain for security when computers happened to be involved, much as politics, health, or finance are. I lean toward security as a big tent, where many disciplines form key parts of the expansive area of cybersecurity. This philosophy led me to define the area of human-centered security in 1996[1] as very much a part of the core of creating secure systems. Today, disinformation is moving beyond usability and social media conferences, like the Association for Computing Machinery (ACM) CHI, ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), and Association for the Advancement of Artificial Intelligence International Conference on Web and Social Media (ICWSM), and starting to appear in security conferences such as the IEEE Symposium on Security and Privacy,[2] signaling broader agreement that disinformation research is part of cybersecurity research.

In that same timeframe, at the Annual Computer Security Applications Conference (ACSAC) in 2019, I was invited to be on a panel on "Disinformation and Harmful Messaging," bringing lessons learned from more than 20 years of usable security to the then-emerging area of combatting disinformation. This editorial looks back at those early insights and adds some current reflections in the light of the state of disinformation research, particularly as it relates to cybersecurity.

My first lesson in 2019 was about misinformation. Disinformation is (in theory) distinct from misinformation since disinformation involves malicious intent, while misinformation does not. I personally find these definitions slippery since misinformation can be used as disinformation. I'm a bit unclear on the converse, whether disinformation relayed without malicious intent is misinformation.

When I was young, I noticed that a close family member got their news from *Weekly World News* ("The World's Only Reliable News"[10]). With cover headlines spanning "Bat Child Found in Cave!" to (after transitioning from paper to an online website) "Obama Appoints Martian Ambassador!", "Flying Cats Terrorize West Virginia!", and "Wife Shrinks Cheating Husband!", the tabloid published sensationalized, mostly fictional "news." I could never tell what my relative thought of this news source; they did not treat it as a joke, nor did they go out of their way to proclaim its truth. I'm sorry now that I never thought to ask about their opinion.

The lesson on misinformation was that people will not only go out of their way to read it but will pay money to do so (at least, they did back when it was in paper form). This shows that, in at least some cases, simply identifying misinformation would not be a sufficient basis for countering disinformation. Facebook seems to have learned that early on, switching from identifying misinformation with a "disputed" flag. It transitioned from that simple identification of disputed information to a pointer to related articles. These related articles were meant to debunk or counter the disputed posts. Presumably, the "disputed" flags alone did not change the Facebook click-throughs in the way they were meant to.

The "disputed" flag was meant to be a disinformation warning, and there has been much research on the effectiveness of security warnings in the usable security area. Research on Transport Layer Security (TLS) website authentication warnings showed that the majority of users ignored them.[3,4] At the time, most technical errors were either intentional (server self-signed certificates) or common human administration errors

Digital Object Identifier 10.1109/MSEC.2022.3159405
Date of current version: 26 May 2022

(allowing server certificate expiration). These false-positive security warnings caused habituation; users got used to them being wrong and clicked on through.

Later warning research looked at phishing warnings, a different class expected to have a substantially lower false-positive rate. This increased likelihood of a true positive allowed browser vendors to design those as "active warnings"—interstitials that interrupt the user's task. They also gave the user choices, made recommendations, and failed safely. Research found that substantially more participants heeded those compared to the previously passive warnings (about phishing sites) that were easily dismissed.

Disinformation researchers recently built on these usable security warning lessons, examining both contextual and interstitial disinformation warnings in the context of a simulated search task.[5] The interstitial warnings had a strong effect on user behavior. The researchers also looked at any effect that informative warnings have (versus uninformative warnings) and any effect that better conveying a risk of harm might have, but neither showed evidence of changing user behavior in the study. In contrast, research studying user-reported intention in the face of behavioral nudges on accuracy found that participants' reported intention to share is impacted by when they are provided with an accuracy assessment and rationale.[6]

These two studies show a contrast in the research tasks chosen to study the potential impact of disinformation interventions, both using crowd workers. One constructs a cover task to study the actions taken (or not), while the other explicitly asks about how participants would intend to act in a constructed scenario. The contrasting results may be due to contrasting approaches (informative warnings versus accuracy nudges) or the method of study (cover task with action versus a question about future intent).

Looking at security warning research for other potential directions for disinformation warnings (beyond being certain of what you're warning about), early studies in Chrome malware warnings found that users heeded warnings about sites they had not visited but were unpredictable for warnings about sites they had visited. When surveyed, users said that they trust high-reputation sites more than malware warnings. Subsequent work on the visual design of those warnings ensuring that they promoted the safe choice increased their impact (by reducing the percentage of users who made the unsafe choice).

Thus, the specifics of both the visual and content design of disinformation warnings and nudges may change outcomes. In addition, the (perceived) provenance of disinformation and who has been sharing it may also vary the user response to warnings (much like the site being warned about a changed user response to malware warnings). Based on these security warning lessons, perceived provenance (of both disinformation and warnings) and how that might impact the belief in and sharing of disinformation is another potentially fruitful area of study.

Digital Object Identifier 10.1109/MSEC.2021.3133131

A question that is often considered in usable security studies is how such security features or defenses fare when they are the target of an attack. We should look more broadly at the disinformation ecosystem to consider the responses if warnings and nudges (or other forms of moderation of disinformation content) are effective. Current disinformation research is looking at the impact of deplatforming and platform migration.[7]

I suggest that, instead of silencing or moving disinformation, effective warnings might themselves be the subject of an attack or subversion. Research could consider how the trust or assurance in disinformation warnings might be subverted. Extending the analogies from usable security research, potential attacks on effective counterdisinformation measures that should be studied include disinformation campaigns against the warnings themselves, attacks on the data sources or technical processes creating the warnings, and spoofing warnings to produce habituation.

The evolution of antiphishing research and defenses provides us with a good example from usable security that illustrates the following lesson: mature security systems are not designed to rely on consistently omniscient user actions as a single line of defense. Early antiphishing defenses explored ways to help the user detect phishing attacks (transitioning such detection to technical measures). Subsequent spear phishing attacks targeted their audience more precisely, with a different look and feel than broader phishing campaigns, making them harder to detect (and avoid). System design and deployment progressed from the detection of phishing attacks to include additional defenses on authentication beyond the stealable and reusable password. Such authentication checks included checking that the browser instance was previously used, that the request was from a geographically viable IP address, and the number of active sessions (an old measure and a weak signal at best). Technologies, such as security keys, could block account takeover entirely when administered with a policy guarding against fallback method takeovers as well.[11]

What lessons about disinformation might we consider from the phishing defense arc I have outlined? Broadly, while specific defenses against misinformation are currently being individually researched, tested, and tried, a robust defense is likely to require a full-system approach in the near term, with both human-centered and technical defenses, in a layered fashion. Such defenses already are likely in use on social media platforms that have policies against disinformation of certain types.

More specifically, in phishing, credentials are the near-term object but only as a means to the end of another action, such as an account takeover. In disinformation, attention is the near-term object as a means to the end of mind share or influence. It is the social media (or web) technology, including "the algorithm" for determining attention and human affordances that feed into that algorithm, that is being leveraged for that human attention. How all of that produces influence on specific human targets is an area of research the needs to be tied to attention research.

Such full-system research should have greater ecological validity that encompasses specific system features as well as a combination of defenses. Modeling or simulation might enable the testing of alternative combinations of defenses. As with other cybersecurity research, testing attacks directly on disinformation protections should be considered. Would "spear disinformation" generated by humans or artificial intelligence (AI) subvert the protections that involve the human in the loop or other technical defenses identifying misinformation?

The ability to research and test disinformation defenses in a rigorous and repeatable fashion is necessary to these research directions. Such testing in the usable security field took off in the area of passwords when a data set of ecologically valid passwords was leaked and made available to researchers.[12]

Data sets with both ground truth and ecological validity are hard to come by in the area of misinformation. Conferences relying heavily on data-driven AI analysis, such as ICWSM, are encouraging their availability through a track dedicated to such data sets. Platforms such as Twitter make available data that can be used as the ground truth in testing through lists, such as their account takedowns. Another challenge to such testing is constructing tests that involve humans and are ethical,[8] controlled, and suitably realistic.

A challenge specific to disinformation is including the right social (or sociological) context in such tests since many campaigns of interest are targeted at a specific population. As with studies of usable security for information security workers (see our special issue in January/February 2023, call for papers at https://www.computer.org/digital-library/magazines/sp/cfp-usable-security-workers), getting the cultural context and interactions right may impact the utility of the particular hypothesis being tested. While the use of behavioral, psychological, and sociological models holds promise for emulating or simulating responses to disinformation [see, for instance, the Social, Cultural, and Behavioral Modeling conference (SBP-BRiMS)] and its counters, current limitations make this an uncompelling substitute for field experiments.

Coming back around to the (somewhat slippery) definitions

of *misinformation* and *disinformation* I began with, the requirements that will lead to data sets combining ecological validity with ground truth are similarly fluid. The current gold standard is a post hoc ground truth determined from a data set "in the wild." For example, lists from platforms of account takedowns related to a topic (such as medical misinformation in the COVID pandemic) are combined with archived content on that narrative to find communications on that topic, allowing an analysis of the technical or human ability to identify disinformation or influential malicious actors.[9]

More targeted data sets may be used to evaluate the specific technologies used in disinformation, such as detecting bots or deepfakes, and determining what aspects of memes are related to them going viral. While bots and deepfakes are, by definition, "false" or synthetic, disinformation campaigns and memes may include information that is not factually false but meant to shift opinion into channels contradictory to those otherwise promoted by a full and fair consideration of the topic or those that promote malicious actions. They share some overlap with intimidation techniques, such as doxing.

Thus, a synthetic data set with ecological validity for full-feature disinformation testing would, in theory, need to both move a specific type of target audience and respond to potential countermoves in a fashion realistic enough to analyze the impact of those countermoves. Defining and using ecologically valid data sets that can be engaged with through analysis and action by technology and humans is a large research challenge in disinformation research, requiring skills from a range of disciplines.

Much as with usable security in 2005 (when the Symposium on Usable Privacy and Security began), I believe that the very diverse community of researchers involved in disinformation, security, usability, sociology, and policy would benefit from a space to call their own, come together, share research on the more cross-cutting aspects of their work (which can be a difficult fit in conferences dedicated to a single discipline), and encourage work on the methodology aspects that need the most progress. In the meantime, those of us involved in conference, journal, and magazine reviews of submissions, such as this one, should welcome such cross-cutting work to increase its reach and impact. ■

Acknowledgment

Distribution Statement A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the U.S. Air Force under Air Force Contract FA8702-15-D-0001. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

References

1. M. E. Zurko and R. T. Simon, "User-centered security," in *Proc. 1996 Workshop New Security Paradigms*, pp. 27–33.
2. M. H. Saeed, S. Ali, J. Blackburn, E. D. Cristofaro, S. Zannettou, and S. Gianluca, "TROLLMAGNIFIER: Detecting state-sponsored troll accounts on reddit," 2021, arXiv: 2112.00443.
3. J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor, "Crying wolf: An empirical study of SSL warning effectiveness," in *Proc. 18th Conf. USENIX Security Symp.*, 2009, doi: 10.5555/1855768.1855793.
4. D. Akhawe, B. Amann, M. Vallentin, and R. Sommer, "Here's my cert, so trust me, maybe?: Understanding TLS errors on the web," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 59–70, doi: 10.1145/2488388.2488395.
5. B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. Nathan Matias, and J. Mayer, "Adapting security warnings to counter online disinformation," in *Proc. 30th {USENIX} Security Symp. ({USENIX} Security 21)*, 2021, pp. 1163–1180.
6. F. Jahanbakhsh, A. X. Zhang, A. J. Berinsky, G. Pennycook, D. G. Rand, and D. R. Karger, "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media," *Proc. ACM Human-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–42, 2021, doi: 10.1145/3449092.
7. S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter," *Proc. ACM Human-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–30, 2021, doi: 10.1145/3479525.
8. E. Niedbala, K. J. Ferguson-Walter, and D. LaFon, "Responsible integration of behavioral science in computer science research and development," in *Proc. IEEE Hawaii Int. Conf. Syst. Sci. (HICSS), Cyber Deception Cyberpsychol. Defense Minitrack*, 2022, pp. 1–10, doi: 10.24251/HICSS.2022.278.
9. S. T. Smith, E. K. Kao, E. D. Mackin, D. C. Shah, O. Simek, and D. B. Rubin, "Automatic detection of influential actors in disinformation networks," *Proc. Nat. Acad. Sci.*, vol. 118, no. 4, p. e2011216118, 2021, doi: 10.1073/pnas.2011216118.
10. "Weekly World News," Wikipedia. https://en.wikipedia.org/wiki/Weekly_World_News (Accessed: Mar. 30, 2022).
11. A. Gaynor, "Quantifying memory unsafety and reactions to it," ENIGMA.usenix.org. https://www.usenix.org/sites/default/files/conference/protected-files/enigma2021_slides_gaynor.pdf (Accessed: Mar. 30, 2022).
12. "RockYou," Wikipedia. https://en.wikipedia.org/wiki/RockYou (Accessed: Mar. 30, 2022).