ACIG
APPLIED
CYBERSECURITY
& INTERNET
GOVERNANCE

NASK

The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats

Mary Ellen Zurko | MIT Lincoln Laboratory, United States |
ORCID: 0000-0002-9427-1607

Received: 09.12.2024

Accepted: 10.12.2024

Published: 27.12.2024

Cite this article as:
Zurko, M. E. " The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats," ACIG, vol. 4, no. 2, 2024, DOI: doi: 10.60097/ACIG/199341

Corresponding author:
Mary Ellen Zurko, MIT Lincoln Laboratory, United States; E-mail: maryellen.zurko@ll.mit.edu
0000-0002-9427-1607

Prof. Mary Ellen Zurko was interviewed by Prof. Aleksandra Gasztold online on 4 December 2024

Drawing from your extensive experience in usable security, what lessons can be applied to the fight against disinformation?

For starters, thank you for inviting me to share my insights on the future of security empowerment and evolving methodologies to counter those threats. While I'm a technologist with CS degrees from MIT, how humans interact with technology has been a continuing interest of mine, going back to my bachelor's thesis (I won't name the year!), and my first job in cybersecurity, where I owned the UI (because no one else wanted to).

I've been active in the usable security research community since defining the area in 1996 (no one had put a name to it, though several others were doing it). One observation from my work in cybersecurity and usable security is that the same patterns and lessons recur (although it is hard to predict WHICH will recur when things change).

Also, I tend to be careful about using the terms disinformation, misinformation, and mal-information. They have different definitions, at least in the research community, that have to do with the intention of the source and sender. Misinformation is the term that assumes the fewest ill intentions, and again, because I'm a technologist, I use that one when talking in generalities that do not involve known ill intentions on the part of all sources.

So, actually back to your question. I would say most generally one of the lessons from usable security is that technologists often struggle to predict how humans will react to new technologies, including misinformation. Early research in usable security found pretty rapidly that we had to actually test or otherwise measure how people would respond to new technology, its new uses, and the threats it imposes on them. Not entirely realistic, very controlled in-lab testing would yield responses different from measuring what people do "in the field" (that's what we call "in real life").

How can effective warnings be designed to combat disinformation, and what lessons can be drawn from Facebook's approach to addressing misinformation and user engagement?

So I want to say my first lesson on misinformation was well before college (again, not naming the decade), when I saw an elderly relative regularly reading a newspaper I had never heard of, called the Weekly World News. This was a tabloid of mostly fictional "news", whose most memorable headline was "Bat Child Found in Cave!". The stories were all largely impossible, but explained in terms that made them sound both plausible and quite sensational. She paid money for that newspaper, and we were not people who had a lot of money for unnecessary items. Thus my first lesson, people will actually go out of their way and pay good money to consume misinformation.

Facebook's initial response to misinformation was to identify it. That didn't make Nana avoid it, and it didn't do much for the Facebook population either. It's said that in some cases it attracted people instead. And again, it certainly attracted my Nana. The usable security community has a very long background in researching how users respond to security warnings. I would say one of the takeaways from that is that if the security warnings themselves are equivocal, if they can't be certain and clear about the harm, if it's only the vague possibility of harm, users will click through them, at an increasing rate as they get accustomed to seeing them time and again. So Facebook changed their initial approach from a vague warning, to a "disputed" flag, with pointers to related articles, which countered or debunked the misinformation [1, p. 4]. I'd also say they have the luxury to test "in the field" and at scale, and that was what they were doing.

In your work, you emphasize the importance of layered defenses, which involve applying a combination of strategies to create a comprehensive system for countering information manipulation. Could you elaborate on this approach, including its technical, educational and warning layers?

Yes, thank you. I'm not only a researcher; I've worked on product. I was security architect for one of IBM's first cloud products. One critical lesson I learned is that securing a system requires viewing it holistically, as a system with all kinds of layered defenses. Layering technical defenses, called "defense in depth", is considered best practice. When humans are involved, making choices and getting things done, then those layered defenses need to include the human, but not get in the way of what they are doing. The evolution of anti-phishing defenses is a great example. Technology alone can't be sure that an email is phishing (or worse still, targeted spear phishing). Various technical responses are unsatisfactory alone, since they can't be sure. Even with anti-phishing education, both the technology and the human can be tricked by ever evolving attacks. Some percentage of users will fall for a strong targeted phishing attack, even when technology, education, and warnings have done their level best (in part because they were also doing their level best on email that wasn't phishing). So the system needs to be designed not only to defend against threats but also to anticipate and mitigate the impact of inevitable breaches and breakdowns.

You and your team have developed the CIOTER system, which integrates these principles into a robust and scalable testbed.

Yes, thank you. Sorry to interrupt you mid point! I have a passionate belief in the importance of testing, in both cybersecurity and usable security. So the goal of developing a testbed for Countering Influence Operations (the CIO in 'seaotter'), by testing the technology involved and the human use of that technology is an exciting one for me.

Can you explain the CIOTER system that you and your team developed. Can you present its purpose, design principles, and potential applications in evaluating and advancing tools for information operations?

I'll try to keep it crisp, but any reader interested in all the wonderful details can read our published paper "A Testbed for Operations in the Information Environment" [2].

CIOTER focuses on building testbed capabilities for assessing technology used in Operations in the Information Environment; technology used to detect and counter misinformation and its cousins. It is inspired in part by cybersecurity testbeds, which are used extensively in education, technology training, and exercises in cybersecurity skills. While cybersecurity testbeds largely focus on network and host based attacks and defenses, our OIE focus is on testbed capabilities that focus on human-readable data and content, services like social media, and how human operators can work with tools to detect and counter misinformation.

We've designed our capabilities to be reusable, redeployable, and reconfigurable, so that they can be used in a variety of contexts, and can interoperate with and complement cybersecurity testbeds.

How does CIOTER's modular architecture facilitate the integration of emerging technologies or adaptation to new adversarial tactics in information operations?

From a technical infrastructure perspective, CIOTER's modularity is achieved through containerization, allowing mix and match with different technologies that process content that might include misinformation in any format; text, memes, videos. A significant focus of CIOTER is on the content pipeline, which not only processes information but also archives and curates different datasets that can represent different adversarial tactics and technologies over time. We can even iteratively test technologies that generate and detect technical changes in content, such as modifications using different forms of AI or ML [3].

In the context of combating disinformation, how does CIOTER contribute to the development of tools like deepfake detection or authorship verification systems?

Both deepfake detection and authorship verification are fairly mature uses of AI technology to detect misinformation. There are curated datasets available, and competitions with established metrics for how well a piece of technology does over a specific dataset. CIOTER can be used to try out a new technology, or an established technology over a dataset modified with a new or different approach. Our extensible metrics engine has all the accepted metrics for success of AI on these tasks, and can be modified with new ones that are tuned to different tradeoffs in things like false warnings. For example, we compared the performance of a specific deepfake detection approach over a corpus that included AI generated deepfakes, and another dataset representing a different threat model; manually modified images (sometimes called "cheap fakes").

Which aspects of disinformation are most easily analyzed using CIOTER? Does the system allow for the evaluation of the effectiveness of counter-disinformation campaigns?

We've got a cool "Over The Shoulder" capability that lets training organizers see how learners and operators are using technology for countering disinformation and other forms of adversarial content. It records all the interactions for viewing during training, and analysis after the event. If a student is confused, or something goes wrong with the tool or how they used it, instructors can can replay the session to pinpoint the issue and help. If there was a 'right' answer and the learner didn't identify it, graders can use the recording to give partial credit if the right keywords appeared, for example, by searching for them. CIOTER also includes dashboards that can show all kinds of activity during an event, for one participant, or a team. The measurements can be correlated with demographic information, so you can look at how different experience levels or roles influence tool use and task completion.

Given the rapid evolution of social media platforms and adversarial techniques, how does CIOTER remain agile and relevant in addressing new threats?

One thing we all know is that social media platforms will come and go, evolve and change. The specific features at a point in time on a social media platform will mean different things at different times (like the blue check on Twitter accounts), and "the algorithm", which determines what each individual sees, will change and effect the impact of both adversarial content and counter disinformation content. To address this, CIOTER is designed to remain agile by incorporating a capability that can flexibly emulate a specific social media platform at a specific point in time, to allow for replay of curated datasets and generated content that reflects what it looks like in various platforms, under different, configurable assumptions.

What do you see as the most critical areas for future research in countering influence operations? Are there specific technological or interdisciplinary advancements you believe are essential to developing more effective defenses against disinformation?

One lesson I learned from pioneering usable security is how challenging it can be to publish research that crosses established boundaries in existing conferences and journals. Rising PhDs and professors need to get their research published, so need to work in areas that are publishable. I've heard professors say that their research on countering influence operations can suffer from this problem; a cybersecurity venue might think it's sociological research, and a sociological venue might point back to cybersecurity publishing opportunities. Just focusing on cybersecurity problems, I'm on a National Academies study of Cyber Hard Problems, and recorded public testimony available on the website includes discussion of how many cybersecurity problems today go beyond just technical problems.

Defending against disinformation and mal-information can involve not just cybersecurity and sociology, but psychology and even political science. There aren't a lot of venues that have specialist reviewers in all those areas. Fostering the best research in countering malign influence operations will require building those communities and venues, that support interdisciplinary work.

* * *

Mary Ellen Zurko is a technical staff member in the Cyber Operations and Analysis Technology Group at MIT Lincoln Laboratory. With over 35 years of experience in cybersecurity and more than 20 patents, she defined the field of user-centered security in 1996. Zurko has worked in research, product development, and early prototyping, and was the security architect of one of IBM's first cloud products. She is a founding member of the National Academies' Forum on Cyber Resilience and serves as a Distinguished Expert for the National Security Agency's Best Scientific Cybersecurity Research Paper competition. Her areas of research focus on unusable security for attackers, zero trust architectures for government systems, security development and code security, authorization policies, high-assurance virtual machine monitors, the web, and public key infrastructure.

References

[1] M. E. Zurko, "Disinformation and Reflections From Usable Security," IEEE Security & Privacy, vol. 20, no. 3, pp. 4–7, 2022, doi: 10.1109/MSEC.2022.3159405.

[2] A. Tse, S. Vattam, V. Ercolani, D. Stetson, M.E. Zurko, "A Testbed for Operations in the Information Environment," CSET '24: Proceedings of the 17th Cyber Security Experimentation and Test Workshop, pp. 83–90, 2024, doi: 10.1145/3675741.3675751.

[3] M. E. Zurko, J. Haney, "Usable Security and Privacy for Security and Privacy Workers," IEEE Security & Privacy, vol. 21, no. 1, pp. 8–10, 2022, doi: 10.1109/MSEC.2022.3221855.