Disinformation 2.0 in the Age of AI: A Cybersecurity Perspective

Why disinformation is a cyber threat.

According to a report from Lloyd’s Register Foundation, at present, cybercrime is one of the biggest concerns of Internet users worldwide, with disinformation ranking highest among such risks (57% of Internet users across all parts of the world, socioeconomic groups, and all ages). For years, there has been a discussion in the security community about whether disinformation should be considered a cyber threat. However, recent worldwide phenomena (for example, an increase in the frequency and sophistication of cyberattacks, the 2016 U.S. election interference, the Russian invasion in Ukraine, the COVID-19 pandemic, and so forth) have made disinformation one of the most potent cybersecurity threats for businesses, governments, the media, and society as a whole. In addition, recent breakthroughs in AI have further enabled the creation of highly realistic fake content at scale. As such, we argue that disinformation should be rightfully considered a cyber threat, and therefore developing effective countermeasures is critically necessary.

The way disinformation is evolving is not exactly new, as other “classical” cyber threats followed a similar path. First, disinformation has been around for centuries, and the Internet is just the latest means of communication used to spread it. We have already witnessed similar developments before. That is, there were different types of crimes such as scams, extortions, and thefts, and we now see their cyber versions, which are less risky for attackers yet more effective than their classical forms. Thus, the use of the Internet (especially social media) made it possible to boost the scale and range at which these attacks can be launched, but the essence of the attack itself remains the same. Similarly, disinformation can affect many more people in a much shorter time than in the case of the non-digital version (for example, traditional newspapers, TV news). Moreover, advances in AI allowed the creation of deepfakes in various types of digital media (that is, images, video, speech) and text, and the introduced modifications are difficult to spot, distinguish, and explain. This greatly enhances the resulting disinformation’s potential reach and believability.
Note that disinformation is not necessarily expected to provide direct revenue, as in the case of other cyber threats. However, such cases already have happened, for example, by spreading disinformation to manipulate stock price\(^e\) or earning income by disseminating it.\(^f\)

**Disinformation 2.0**

In the conventional paradigm of disinformation, on the attack side, we have disinformation creators who fabricate false information and post it to websites or social media for various purposes such as monetary incentives or political agenda. On the defense side, platforms have used human operators as well as computational methods to ensure the integrity of information, such as disinformation detectors to filter out questionable content, and socialbot detectors to curb the spread of disinformation. By and large, so far, creators (semi-) manually have created and disseminated disinformation content without using sophisticated AI techniques. When fake content is detected and filtered out by defense mechanisms, creators would simply attempt to redisseminate it using different accounts or domains.

With the recent advances in AI, we envision this paradigm is likely to change substantially, yielding what we call *disinformation 2.0*, where disinformation would become more targeted and personalized, its content indistinguishable from genuine content, and its creation and dissemination further accelerated by AI. Disinformation 2.0 will increase distrust in news that humans encounter in real and digital worlds, which is a major problem already.\(^g\) Typically, in cyberspace, an attacker’s aim is to find weak spots in the targeted system and exploit/compromise them using technical and/or non-technical means. If we treat disinformation as a cyberattack, then several scenarios of disinformation 2.0 become possible, as illustrated in Figure 1.

1. Adversaries generate more convincing disinformation, using generative AI techniques (for example, ChatGPT for texts or DALL-E for images);\(^h\)
2. Adversaries subtly perturb existing content using AI methods to create more convincing disinformation with better capability to evade detection;\(^i\)
3. Adversaries use AI methods to attack disinformation detection ecosystems, promoting disinformation and demoting real news;\(^j\) and
4. Adversaries strategize the spread of disinformation using AI methods, maximizing its influence on the (social) network while evading socialbot detection ecosystems.\(^k\)

When the creation and dissemination of disinformation are greatly enhanced using advanced AI methods in one of these scenarios, the resulting disinformation 2.0 becomes much more difficult to detect and more persuasive/impactful than the previous disinformation 1.0.

**Countermeasures**

Detecting disinformation 1.0 has been extensively researched in recent years (for example, see Kim et al.\(^l\) for a survey), and many solutions have reported high detection accuracies (for example, Cui et al.).\(^m\) However, there are still several remaining issues, including early detection, multilingual/multiplatform detection, better explainability, or sociotechnical issues.\(^n\) As with every cyber threat, completely eliminating disinformation is unlikely (as achieving complete security is never possible). We must diminish the impact of disinformation on Internet users, as we did with threats such as spam email messages. Several years ago, for instance, spam email messages were considered one of the major threats, but now their scale and relevancy are not as high as they were before.\(^o\) This has been achieved due to decades of research advances, during which many sophisticated techniques resulted in significantly limiting the volume of spam email messages in Internet users’ inboxes.

Currently, a major disadvantage of our defenses against disinformation 2.0 is that they are being individually researched, developed, deployed, and evaluated, which is not very effective in diminishing the threat of disinformation.
The problem is likely to require a layered approach, where both human-oriented (for example, raising awareness, training, and so forth) and technical measures are applied at the news creation, transmission, and consumption layers. We propose to distinguish four layers at which disinformation impact can be diminished.

1. **Social Network-level layer** is organized by the social network operator.

2. **ISP-level**
   - Whitelist/blacklist of known websites hosting disinformation
   - Beating filter bubble
   - Fact-checking websites
   - News author reputation verification
   - Raising awareness training
   - Responsible news sharing

3. **Device-level**
   - Disinformation detection and spread control
   - Bots detection
   - News source verification
   - Whitelist/blacklist of news sources
   - Improved information suggestions algorithms

4. **User-level**
   - Automatic disinformation detection (AI-based)
   - In-browser news source verification (for example, via watermark)
   - Whitelist/blacklist of news sources
   - Enforcing user’s responsible news sharing
   - Automatic cross-verification of the news at different sources (AI-based)

For this challenge, we argue to use some lessons learned from cybersecurity, where typically multiple “layers of defense” are envisioned. On the Internet, as security risks occur at various levels, it is necessary to set up security mechanisms that provide multiple layers of defense against these risks (for example, by considering threats at the system, network, application, and transmission levels).

A popular approach to layered security is defense-in-depth architectures, where controls are designed to protect the physical (that is, prevent physical access to IT systems, such as security guards or locked doors), technical (that is, security measures that use specialized hardware or software, for example, as a firewall, IDS/IPS, antivirus), and administrative (that is, policies or procedures directed at an organization’s employees) aspects of the communication network. Using such a layered approach to provide network security makes it possible for an attacker who penetrates one layer of defense to be stopped by a subsequent layer. Therefore, addressing the problem is likely to require a layered approach, where both human-oriented (for example, raising awareness, training, and so forth) and technical measures are applied at the news creation, transmission, and consumption layers. We propose to distinguish four layers at which disinformation impact can be diminished (see Figure 2).

1. **Social Network-level layer** is organized by the social network operator.

The way disinformation is evolving is not exactly new, as other “classical” cyber threats followed a similar path.
To be effective, all security mechanisms and user actions must be applied in tandem.

Some of the solutions in the proposed approach may be considered invasive from a user-privacy point of view, and protecting user data privacy is a critical concern in today's digital landscape. Fortunately, several effective solutions and strategies already exist that can be employed to fix this issue and follow the privacy-by-design principle, for example, by incorporating schemes relying on differential privacy, secure data aggregation, homomorphic encryption, or data masking and tokenization.

We strongly believe the advanced AI techniques, despite their benefits to society, greatly enable adversaries to achieve more sophisticated and effective disinformation 2.0 attacks. As such, adopting the lessons learned from cybersecurity research, novel countermeasures are needed, especially a holistic layered approach as discussed.

References
10. Zurko, M.E. Disinformation and reflections from usable security. IEEE Security and Privacy 20, 3 (2022); 10.1109/MSEC.2022.3199406

Wojciech Mazurczyk (wojciech.mazurczyk@pw.edu.pl) is a professor at the Institute of Computer Science at Warsaw University of Technology, Warsaw, Poland.

Dongwon Lee (dongwon@psu.edu) is a professor in the College of Information Sciences and Technology at The Pennsylvania State University, USA, PA.

Andreas Vlachos (av308@cam.ac.uk) is a professor of Natural Language Processing and Natural Language Processing in the Department of Computer Science and Technology at the University of Cambridge, U.K.

Wojciech Mazurczyk acknowledges the funding obtained from the EIG CONCERT-Japan call to the project Detection of disinformation on Social Media Platform Rms ‘‘SIMILAR’’ through grant EIG CONCERT-JAPAN/05/2021 (National Centre for Research and Development, Poland). Dongwon Lee was in part supported by NSF awards #1820609, #2114824, and #2113144. Andreas Vlachos acknowledges the funding from ERC grant AVERITEC (GA 865958), and the E.U. H2020 grant MONITIO (GA 865576).