WASP—HS

AI and Cyber Security

Report 2024

Table of Contents

Wallenberg AI, Autonomous Systems and Software Program - Humanity and Society (WASP-HS) would like to thank all chairs and participants of the event AI and Cyber Security for contributing to the fruitful discussions which this report is based on.

Introduction

The value of information, and the problem of withholding valuable information was at the core in a famous exchange at the first Hacker's Conference in 1984 between Stewart Brand and Steve Wozniak. Brand agreed that the value of holding the right information at the right time is unquestionable. However, Brand said, "information almost wants to be free because the cost of getting it out, in many respects, is getting lower and lower all the time".

Thus, Brand settled two important facts; first, that information is a highly valuable asset, for individuals and organizations alike, and second, that information is difficult to protect. 40 years later, we live digital lives and our personal, professional, and business lives have come to depend almost completely on the functionality and reliability of IT- systems. As we reap the benefits of this development, our dependence on information technology has also increased our vulnerability.

IBM defines a cyberattack as the "intentional effort to steal, expose, alter, disable, or destroy data, applications, or other assets through unauthorized access to a network, computer system or digital device". Cyber security then, is the art of reducing the risk of cyberattacks. Or perhaps more correctly it is the art of reducing the cost of a cyberattack, because experts agree, and experience suggests, that cyberattacks has become a defining characteristic of our digital lives. For instance, The Swedish Civil Contingencies Agency reports a significant increase in cyberattacks on Swedish authorities in 2023, stating that cyberattacks have become the most common type of cyber incidents surpassing system errors and mistakes.

Cyberattacks are used successfully not only to steal valuable information but to disturb business operations and stir insecurity and uncertainly in organizations and societies. The tense geopolitical climate, and growing polarization of national and global political discourse, appears to provide particularly fertile grounds for trying out new tools in the arsenals of malicious actors. AI technology make such weapons even more potent.

This WASP-HS Community Reference Meeting was organized around two keynote talks with two roundtable discussions in between. Our opening speaker Staffan Truvé is a member of the board of our sister program WASP, and Chief Technical Officer and co-founder of Recorded Future. Truvé's keynote focused on a service developed to detect data leaks and assess and monitor threats through a combination of AI technologies and big data analytics. Our closing speaker Charles Berret is a postdoc in critical data visualization at Linköping University. His keynote presented The Drawbridge Model, which provides an analytical tool to categorize forms of information security based on different conditions of success and failure in communication.

The two roundtables were chaired by Francis Lee from Chalmers University and Helena Lindgren from Umeå University, both members of the WASP-HS management team. The roundtable discussions touched upon a range of issues related to threats such as sleeper agents in AI systems and data poisoning, and protective strategies such as training operators with simulated cyberattacks.

WASP-HS aims not only to contribute knowledge to the development of innovative technologies for increased cyber security. Our chief concern is the human and social context of cyber security, and we aim to also consider potential collateral damage for individuals, companies, and society in the arms race between the cyber-attacker and defender. At what stage does increased cybersecurity interfere with personal integrity? And does too much attention to security risk negatively affecting our trust in technology, in the state, in society, in fellow citizens? It is obvious that individuals, organizations and states all have work to do. Awareness, training, and technological development is equally important to strengthen cybersecurity. And the approach has to be intellectually broad, because at the core, there are some serious social values at stake.

Christofer Edling, WASP-HS Program Director

WASP-HS Community Reference Meetings (CRMs) are meeting places for Swedish private and public organizations and WASP-HS researchers. Each meeting has a specially selected theme with the aim of bringing business and research together to expand knowledge and strengthen collaboration.

This report is based on the discussions and conclusions from the CRM on the topic of AI and cyber security. The event took place on 26 March, 2024.

Highlights From Keynote
AI for Cyber Security and How Analytics Can Prevent Cyber-Attacks

Keynote Speaker
Staffan Truvé, CTO and co-founder of Recorded Future AB, Member of the board of WASP

Staffan Truvé gave a keynote presentation that focused on how large-scale data collection from the internet and AI based analytics can help predict and prevent cyber-attacks.

He presented how Recorded Future is developing a hybrid intelligence system combining different artificial intelligence techniques with human analysts' expertise to combat cyber security threats and influence operations. The major challenges lie in the vast amount of data and the global shortage of cyber security experts. Fighting modern cyber-attacks is a big-data analytics problem and another problem is that the attackers have access to the same developments in technology, which results in a form of "arms race." It is necessary to monitor what possible attackers are doing in real time. Therefore, Recorded Future builds a threat-oriented digital twin of the world, with machines and humans working together.

Recorded Future uses massive data extracted from the internet in order to build the digital twin of the world. This data is organised in a knowledge graph that captures facts and events and the relationships between them. Software agents are created that subscribe to changes in the graph and calculate risk scores. Explanations are given, based on logics-based reasoning. Large language models (LLMs) augmented with the knowledge graph content are embedded to respond to queries by the analyst and to provide summaries. There is an element of speculation used by analysts when they generate cases to assess.

There are a number of research and development challenges with machines and humans working together (human-AI teaming) to combat cyber threats that need to be addressed. Truvé mentioned dialogue management and intention recognition; the difficulty of language models identifying important information; higher-order reasoning; how to combine visual analytics with language. Further, the detection of deepfake and influence operations is challenging, as is developing ways in which multiple humans communicate with multiple LLM systems. To summarise, the "arms race" is continuing, motivating efforts to improve work routines and to invest in multi-disciplinary research and development of more advanced and efficient ways in how humans and machines can work together to monitor and prevent cyber-attacks. sion, or apparent nonsense. Types of mundane failure at the level of meaning include deception, misunderstanding, and false attribution. Cybersecurity threats at the level of meaning include social engineering, spoofing, and phishing.

Because the Drawbridge Model is rooted in sociocultural analysis, this framework can be highly flexible and medium-agnostic, explaining how historical forms of information security work as well as offering ways to theorize emerging forms of information security.

The model has direct implications for AI and cyber security because it enables us to recognize how different forms of information security leverage different kinds of communication failure to selectively divide an audience. Modeling threats and vulnerabilities using the Drawbridge Model can help develop 'defense in depth' by evaluating the strength of security at each stage. Identifying the ways AI succeeds and fails at communication can point to opportunities for building better defenses. Specifically, generative pre-trained transformers (GPTs) for text will pose a considerable threat at the level of meaning, where previously only human beings could conduct viable social engineering to exploit systems by manipulating other human actors. Attending to the different layers of this model thus offers explanatory and diagnostic affordances for protecting against emerging cyber security threats from AI.

Highlights From Keynote

The Drawbridge Model of Cryptographic Communication: A Framework for Sociocultural Analysis of Information Security

Keynote Speaker
Charles Berret, Postdoctoral fellow in critical data visualization at Linköping University

The Drawbridge Model is a multistage framework to categorize and analyze forms of information security based on different conditions of success and failure in communication.

What can we learn about concealing communications by studying different forms of communication failure? Let's start with a broad, sociocultural definition of cryptography: Any form of communication that uses known sources of communication failure to create reversible, non-destructive encodings that selectively limit the audience of a message. By extension, information security techniques work by selectively raising and lowering the drawbridge of communication's success or failure.

Whether it's Pretty Good Privacy (PGP), invisible ink, or quantum cryptography, information security has always been a matter of designing some means of selecting who gets signal and who gets only noise. Different forms of information security thus serve as means of selectively mediating conditions of communication failure in distinct cases, including those that have or will be impacted by AI.

The Drawbridge Model presents a chain of islands separated by gaps that represent potential sources of communication failure. Successful communication must bridge each gap, while failure at any stage makes further stages inaccessible. For information security to function, these sources of failure must be selective and reversible (like a drawbridge).

1. Access
This means a barrier or lack of authorization keeps you from receiving a message. If a message is inaccessible, communication fails even if message is perfectly viable otherwise. Security schemes that mediate access include keys, envelopes, logins, and file permissions.

2. Recognition
Given access, recognition is the awareness that a specific message or information is present. Communication fails when we do not recognize that the message is there. The message could be perfectly viable otherwise, but communication still fails if unrecognized. Security schemes that mediate access include steganography, invisible ink, and digital watermarks.

Charles Berret

3. Legibility
This model uses a specific definition of legibility drawn from the field of typography. A legible message is rendered using individual symbols known to the user or understood by a machine. The principal security scheme mediating access is CAPTCHA.

4. Readability
In contrast to legibility, readability is the capacity to recognize coherent patterns, words, and syntax in a set of known symbols. Types of mundane readability failure include spelling and encoding errors, incompatible file formats, and bit rot. Security schemes that mediate readability include anagrams, acrostics, and digital encryption.

5. Meaning
Successful transmission of meaning is shared intersubjective accord on the content expressed in a message. Failure is mistaken apprehension, confusion, or apparent nonsense. Types of mundane failure at the level of meaning include deception,

misunderstanding, and false attribution. Cyber security threats at the level of meaning include social engineering, spoofing, and phishing.

Because the Drawbridge Model is rooted in sociocultural analysis, this framework can be highly flexible and medium-agnostic, explaining how historical forms of information security work as well as offering ways to theorize emerging forms of information security.

The model has direct implications for AI and cybersecurity because it enables us to recognize how different forms of information security leverage different kinds of communication failure to selectively divide an audience. Modeling threats and vulnerabilities using the Drawbridge Model can help develop 'defense in depth' by evaluating the strength of security at each stage. Identifying the ways AI succeeds and fails at communication can point to opportunities for building better defenses. Specifically, generative pre-trained transformers (GPTs) for text will pose a considerable threat at the level of meaning, where previously only human beings could conduct viable social engineering to exploit systems by manipulating other human actors. Attending to the different layers of this model thus offers explanatory and diagnostic affordances for protecting against emerging cyber security threats from AI.

Roundtable Discussion

Complexities of AI in Cyber Security

Author

Francis Lee, Associate Professor in Science and Technology Studies at Chalmers University of Technology

Main Challenges

- Sleeper agents and backdoors in AI systems
- Data poisoning and poisoned prompts
- Perceived neutrality can be exploited for nefarious purposes
- Embedded systems in products needs to be protected
- Trade-offs between privacy and security

During the discussions, a significant focus was placed on the challenges posed by sleeper agents in AI systems. A detailed report revealed the difficulties in detecting and eliminating backdoor behaviors in large language models. It was noted that even with knowledge of such vulnerabilities, eradicating them entirely is problematic, thereby underscoring the importance of using comprehensive validation datasets to ensure model integrity.

The meeting also addressed the issue of data poisoning, highlighting the critical need for using clean, reliable data in AI systems to prevent the induction of unintended behaviors. The analogy of "poisoned prompts" was drawn with dog whistling in media studies, suggesting that certain terms can trigger specific, often undesirable, responses among targeted audiences, reflecting challenges in managing the heterogeneity of language understanding by AI.

Participants discussed the perceived neutrality of AI systems and the potential for their exploitation. The possibility of developing technological solutions to protect against such exploits, particularly in data or image manipulation aimed at misleading AI systems, was explored.

The broader role of AI in society was a key part of the agenda, with discussions on the strategic responses needed to mitigate risks, including those from cyber-attacks. The implementation of new digital policies such as the Digital Services Act and Digital Markets Act was recognized as crucial in shaping public and private sector trust and application of AI.

Insights were shared on using AI to enhance product security and development, emphasizing the challenges of protecting smaller AI-driven components from hacking or theft. The session concluded with a reflective discussion on the desired future of AI in society. Questions were raised about AI's potential role as either a tool for societal control or as a beneficial technology, highlighting the moral and ethical considerations that must guide AI development and application.

The diverse viewpoints provided during the meeting offered a rich foundation for ongoing dialogue and development in the field of AI and security, emphasizing the need for a comprehensive understanding of AI's interaction with cybersecurity to effectively navigate future policies and technologies.

Highlights From Keynote
Cyber Security in Practice

Author
Helen Lindgren, Professor in Computer Science at Umeå University

The round-table discussion evolved around how the human is one of the targets, but also how the human is the weakest link. It was pointed out that increase in security typically means decrease in usability. Recommended use of 2-factor authentication and measures such as backing up of data require more engagement and knowledge among people using systems. On the other hand, reports show that Swedes take data protection more seriously now than earlier.

The discussion evolved initially around how AI-based training systems could be used to increase understanding of threats. To be exposed to whole scenarios and simulated situations including both systems and humans could help people increase knowledge by experience situations. Examples were given where large language models (LLMs) were used to try to predict scary futures, including receiving fake voice mails from people's CEOs.

In the professional analysts' perspective who is monitoring systems that are analysing potential threats, it is important to provide tools for how to interpret signals from the system.

Regulatory aspects were discussed briefly. Since companies are by law required to monitor attacks on their systems, the use of systems like the one provided by Recorded Future provides companies the tools to do what is required from them. Monitoring organisations, actors, social networks, and events that can be classified as threats or attacks is necessary to also predict increase in hostile activities. When the question about where the line is drawn for collecting person-related information, facial recognition was considered beyond the line.

Geo-political analyses using multi-lingual AI-based systems were discussed as a means to understand trends. Recorded Future provided examples of using LLMs to analyse how different countries speak about events across different languages, and how rumours are started and spread. It was discussed how detection of such events could give indications of future threats. However, the inherent difficulty in detecting intentions to attack was illustrated by the Hamas attack on Israel, indications of which could not be seen in the massive data collected across public channels in post-event analyses.

How to utilize the hallucination mechanisms of LLMs to simulate possible attacks was discussed and exemplified. Even simulating attacks that seem to be impossible in some interpretation of the world, could be elicited to be possible in a different interpretation of the world.

The round-table conversation can be summarized in the following three key points and directions for interdisciplinary research and development across disciplines including the social sciences and the humanities:

1. Build knowledge about the personal, socio-technical, and cultural aspects of cyber security, and better readiness to prevent attacks.

2. Build training and educative systems to increase knowledge and experience of situations for the general public and for analysts.

3. Build better technology, e.g., improve dialogue systems to recognize intentions, and to act accordingly, such as challenge a person when needed in a situation that requires paying attention to potential threats. Integrating an AI agent within software, for instance, could prompt users to reconsider actions, thereby preventing getting targeted.

This Report

This report is made possible by the Wallenberg AI, Autonomous Systems and Software Program - Humanity and Society (WASP-HS), a national research program in Sweden. The vision behind WASP-HS is to promote new interdisciplinary knowledge in the humanities and social sciences on the subject of artificial intelligence and autonomous systems and their impact on human and social development. The research program enables cutting-edge research, expertise and competence building in the humanities and social sciences. In total, the Wallenberg Foundations are investing up to SEK 660 million in the WASP-HS research program.

For more information please visit www.wasp-hs.org.

For questions or inquiries please contact us at contact@wasp-hs.org.

Authors

Charles Berret, Postdoctoral fellow in critical data visualization at Linköping University

Christofer Edling, Professor of Sociology at Lund University

Helen Lindgren, Professor in Computer Science at Umeå University

Francis Lee, Associate Professor in Science and Technology Studies at Chalmers University of Technology

How to cite this report

WASP-HS. Community Reference Meeting: AI and Cyber Security. Report. June 2024.

WASP—HS

Wallenberg AI, Autonomous Systems and Software Program
- Humanitiy and Society