

Cybersafety Threats – from Deception to Aggression

Edited by

Zinaida Benenson¹, Marianne Junger², Daniela Oliveira³, and Gianluca Stringhini⁴

1 Universität Erlangen-Nürnberg, DE, zinaida.benenson@fau.de

2 University of Twente, NL, m.junger@utwente.nl

3 University of Florida – Gainesville, US, daniela@ece.ufl.edu

4 Boston University, US, gian@bu.edu

Abstract

A number of malicious activities, such as cyberbullying, disinformation, and phishing, are becoming increasingly serious, affecting the wellbeing of Internet users both financially and psychologically. These malicious activities are inherently socio-technical, and therefore effective countermeasures against them must draw not only from engineering and computer science, but also from other disciplines. To discuss these topics and find appropriate countermeasures, we assembled a group of researchers from a number of disciplines such as computer science, criminology, crime science, psychology, and education. Through five days of brainstorming and discussion, the participants developed a roadmap for future research on these topics, along four directions: modelling the attackers, measuring human behavior, detection and prevention approaches for online threats to adolescents, and understanding unintended consequences of mitigation techniques.

Seminar July 21–26, 2019 – <http://www.dagstuhl.de/19302>

2012 ACM Subject Classification Social and professional topics → Computer crime, Security and privacy → Human and societal aspects of security and privacy, Security and privacy → Social engineering attacks

Keywords and phrases Cybersafety, Legal and Ethical Issues on the Web, Online Social Networks, Security and Privacy

Digital Object Identifier 10.4230/DagRep.9.7.117

Edited in cooperation with Matthew Edwards


1 Executive Summary

Zinaida Benenson (Universität Erlangen-Nürnberg, DE)

Marianne Junger (University of Twente, NL)

Daniela Oliveira (University of Florida – Gainesville, US)

Gianluca Stringhini (Boston University, US)

License  Creative Commons BY 3.0 Unported license

© Zinaida Benenson, Marianne Junger, Daniela Oliveira, and Gianluca Stringhini

A number of malicious activities are prospering online and are putting users at risk. In particular, cyber deception and cyber aggression practices are increasing their reach and seriousness, leading to a number of harmful practices such as phishing, disinformation, radicalization, and cyberbullying. Attack strategies include controlling and operating fake or compromised social media accounts, artificially manipulating the reputation of online entities, spreading false information, and manipulating users via psychological principles of influence into performing behaviors that are counter to their best interests and benefit the attackers.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Cybersafety Threats – from Deception to Aggression, *Dagstuhl Reports*, Vol. 9, Issue 7, pp. 117–154

Editors: Zinaida Benenson, Marianne Junger, Daniela Oliveira, and Gianluca Stringhini



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

So far, computer science research on cybersafety has looked at the various sub-problems in isolation, mostly relying on algorithms aimed at threat detection, and without considering the implications of the attacks and countermeasures for individual users as well as for society. On the other hand, human factors and social science researchers often consider user interfaces and social interactions without taking full advantage of the algorithmic, data-driven cybersafety research. Moreover, the legal and ethical implications of attacks and countermeasures are often unclear.

The goal of the Dagstuhl Seminar 19302 “Cybersafety Threats – from Deception to Aggression” was to provide a platform for researchers to look at the problem of cybersafety from a holistic and multi-disciplinary perspective. The participants were drawn from a number of disciplines such as computer science, criminology, psychology, and education, with the aim of developing new ideas to understand and mitigate the problems.

At the beginning of the seminar, we asked participants to identify important themes to focus on, and these themes were refined through specific activities and discussions during the first day: Firstly, all participants gave 5-minute talks where they presented their current research related to the seminar, and their expectations and topics they would like to work on during the week. Secondly, we conducted three *introductory panels* on the topics of *Cyber Deception*, *Cyber Aggression* and *Propaganda & Disinformation*. Each panel consisted of five participants. We took special care to represent different disciplines and different career stages in each panel.

By the beginning of the second day, participants had identified four key themes to study in this area, which we describe in detail in the rest of this section. The participants formed working groups (WGs) for each theme.

Theme 1: Attacker modeling

The working group focused on predicting the next steps of an ongoing attack by means of a probabilistic model. The initial model developed by the group consists of 9 variables: attacker goals, characteristics of the attack (e.g., how long the attack takes, tools employed), consequences, authorization, attribution, expected resilience of the victim, expected characteristics of the victim from attacker’s perspective, actual characteristics of the victim, actual responsiveness of the victim. The developed model was verified and refined using two known attacks as case studies: the Internet Worm (1988) and the SpamHaus DDoS attack (2013).

Two most important next steps to refine the model are:

1. Convert the variables into measurable quantities
2. Obtain labeled data on which the model can be trained

The working group started working on a conceptual paper that describes the model, and discussed possible venues for its publication. Several methods of obtaining the data for the model were proposed, such as interviewing CISOs and other defenders, creating financial incentives for organization to share their data, and organizing a stakeholder workshop including not only defenders, but also former attackers who now work as security consultants.

Theme 2: Unintended consequences of countermeasures

This working group focused on an often overlooked aspect of computer security research: the fact that deploying any countermeasure to mitigate malicious online activity can have unexpected consequences and harms to other parties. The members of this working group started by discussing a number of scenarios: intimate partner abuse, CEO fraud, disinformation, online dating fraud, and phishing, and developed a taxonomy of these potential harms.

The taxonomy takes into account not only technical issues that might arise from deploying countermeasures but also socio-technical ones such as the displacement effect of attackers moving to other victims, the additional costs incurred by using the countermeasure, and the issues arising from complacency, for example leaving users desensitized by displaying too many alerts to prevent a certain type of attack.

Theme 3: Measuring human behavior from information security (and societal) perspectives

Measuring online behavior is of fundamental importance to gain an accurate understanding of malicious online activities such as cybercrime. The research community, however, does not have well established techniques to accurately measure this behavior, and this can lead to studies presenting largely contradicting results. This working group focused on identifying techniques relevant to measure and model various types of online behavior, from cyberbullying and disinformation to ransomware and phishing. As a final outcome, the working group drafted two methodological frameworks for researchers aiming to study these problems, one focused on socio-technical threats (cyberbullying and disinformation) and one focused on cybersecurity (phishing and malware).

Theme 4: Prevention, detection, response and recovery.

A key challenge when mitigating socio-technical issues is developing the most effective countermeasures. This group focused on developing detection and prevention approaches focusing on threats encountered by adolescents when surfing the Web (e.g., cybergrooming). A common issue here is that adolescents rarely turn to adults for help, and therefore any mitigation based on direct parental oversight has limited effectiveness. To go beyond these issues, the group developed a mitigation strategy based on a “guardian angel” approach. The idea is to let a minor create a “guardian avatar” that will then advise them on cybersafety practices, with a decreasing level of oversight as the minor grows up. While the children are very young, the guardian avatar will closely supervise them, reporting any suspicious contacts that they have online to a parent or a guardian. Later, as the child enters adolescence, the avatar will gradually take on an advisory role, eventually only providing advice once the adolescent asks for it. The group considered privacy issues and interdisciplinary aspects related to psychology and education, and developed a proposal of how the avatar would work.

Conclusion and Future Work

The seminar produced a number of ideas on how to investigate and mitigate cybersafety threats. It enabled researchers from different disciplines to connect, and set the agenda for potentially impactful research to be carried out in the next years. Joint publications and funding for joint research were discussed in each WG and later in the plenum. For example, WG 3 considered possibilities for a large international grant, such as H2020. The ideas produced as part of theme 4 resulted in the paper “Identifying Unintended Harms of Cybersecurity Countermeasures” to appear at the APWG eCrime Symposium in November 2019.

2 Table of Contents

Executive Summary

Zinaida Benenson, Marianne Junger, Daniela Oliveira, and Gianluca Stringhini . 117

Overview of Talks


Empirically measuring the economic impact of cyber attacks <i>Abhishta Abhishta</i>	122
Teaching People Not to Fall for Cyber Deception Might Be Harmful <i>Zinaida Benenson</i>	122
Inconsistent Deception and Attribution <i>Matt Bishop</i>	123
Research in Social Engineering <i>Jan-Willem Bullée</i>	124
The Federal Trade Commission <i>Joe Calandrino</i>	124
MITRE's Human Behavior and Cybersecurity Research and Capabilities <i>Deanna Caputo</i>	125
Towards Cognitive Security <i>Claude Castelluccia</i>	126
Measuring Online Radicalisation <i>Yi Ting Chua</i>	126
The Neurobiology of Financial Abuse <i>Natalie Ebner</i>	126
Research in Online Fraud <i>Matthew Edwards</i>	127
The Sociology of Phishing <i>Freya Gassmann</i>	127
Caught in the Crossfire / The language of aggression, violence, and cybercrime <i>Alice Hutchings</i>	128
Psychological aspects of Cybercrime <i>Marianne Junger</i>	129
Research in Security Risk Management <i>Katsiaryna Labunets</i>	130
Research in Phishing <i>Elmer Lastdrager</i>	130
Phishing Susceptibility as a Function of Age, Gender, Weapon of Influence, and Life Domain <i>Daniela Oliveira</i>	130
Cyber Deception and Cyber Aggression <i>Simon Parkin</i>	131

Get to know your geek: towards a sociological understanding of incentives to develop privacy-friendly free and open source software <i>Stefan Schiffner</i>	131
Characterizing Disturbing and Reactionary Content in Youtube <i>Michael Sirivianos</i>	132
Characterization, Detection and Mitigation of Antisocial Behaviour <i>Ivan Srba</i>	132
Measuring and Modeling the Online Information Ecosystem <i>Gianluca Stringhini</i>	133
DISinformation as a Political Game <i>Gareth Tyson</i>	134
Language-based deception detection <i>Sophie van Der Zee</i>	135
Applying Routine Activity Theory to Cybervictimization: A Theoretical and Empirical Approach <i>Sebastian Wachs</i>	135
Research in Evidence-based Security <i>Victoria Wang</i>	136
Deception and deterrence <i>Jeff Yan</i>	137
Working groups	
Theme 1: Attacker Modeling Group <i>Abhishta Abhishta, Zinaida Benenson, Matt Bishop, Joe Calandrino, Natalie Ebner, Manuel Egele, William Robertson, Victoria Wang, and Savvas Zannettou</i>	137
Theme 2: Unexpected Consequences of Countermeasures <i>Matthew Edwards, Yi Ting Chua, Alice Hutchings, Daniela Oliveira, Simon Parkin, Stefan Schiffner, and Gareth Tyson</i>	143
Theme 3: Measuring Human Behavior from Information Security and Societal Perspectives <i>Ivan Srba, Katsiaryna Labunets, and Sophie van Der Zee</i>	150
Theme 4: Prevention, Detection, Response and Recovery <i>Gianluca Stringhini, Freya Gassmann, Marianne Junger, Elmer Lastdrager, Michael Sirivianos, and Sebastian Wachs</i>	152
Participants	154

3 Overview of Talks

3.1 Empirically measuring the economic impact of cyber attacks

Abhishta Abhishta (University of Twente, NL)

License  Creative Commons BY 3.0 Unported license
© Abhishta Abhishta

Measuring the economic impact of cyber crime just by the use of surveys does not provide an accurate picture of the real losses. Well, if you ask people what they don't know, they are bound to provide you with the perception of the answer, which might not be the real answer. This is one of the reasons why we see the losses due to cybercrime being reported in millions of dollars.

A solution for this problem is to empirically measure the economic impact of cyber crimes. This can be done by using many of the newly collected datasets such as the OpenINTEL [1]. However, this method has its own shortcomings. It is not always possible to get the datasets that can be used to measure economic impact (privacy reasons). An example of this is collection of “work study”/“time study” measurements in an IT firm to estimate the true impact of IT downtime due to a cyber attack. “Work study”/“time study” methods have been used in the manufacturing industry to measure the impact of downtime in assembly lines.

The research question I have for this workshop related to the problem described above is: *How can we in a privacy friendly way take “work study” / “time study” measurements at an IT company?*

Another research question that I am interested in and is related to the theme of the workshop is: *As fake news has been around even before the internet, can we learn from how the history has dealt with fake news and use the similar solutions for the current problem.*

References

- 1 Abhishta, A., van Rijswijk-Deij, R., & Nieuwenhuis, L. J. M. *Measuring the impact of a successful DDoS attack on the customer behaviour of managed DNS service providers.* Computer communication review, 48(5), 70-76, 2018

3.2 Teaching People Not to Fall for Cyber Deception Might Be Harmful

Zinaida Benenson (Universität Erlangen-Nürnberg, DE)

License  Creative Commons BY 3.0 Unported license
© Zinaida Benenson

In 2014, my colleagues and I conducted a phishing experiment with a (then) novel design: We recruited over 1200 university students for a study on online behavior, but sent to them a simulated phishing message from a non-existing person. The message referred to a party last week, and contained a suspicious link to the party pictures. After several days, we sent to the participants a questionnaire that debriefed them about the true purpose of the study, and asked them for reasons of their clicking behavior. The most frequently reported reason for clicking was curiosity (34 percent), followed by the explanations that the message fit recipient's expectations (27 percent), as they attended a party last week. Moreover,

16 percent thought that they might know the sender. These results show that decisional heuristics for message processing are relatively easy to misuse, if the attack message refers to work or life interests of the people, or spoofs a known sender.

Defense against spear phishing and other targeted attacks seems to be especially challenging because of the ambiguity of the situations that they create, making the context and content of the message look plausible and legitimate. Because of this ambiguity, asking people to be permanently vigilant when they process their messages might have unintended negative consequences. For example, if their job requires processing a lot of invoices sent via email, they might click on a ransomware-infected file called `invoice.doc`, as this fits their job expectations. But if they are taught to be careful with invoices, they might start missing or delaying the real ones, which stands in a direct conflict with the requirements of their job. Under these circumstances, the employees are likely to disregard this kind of user education attempts after some time, because the only way for them to get their job done in time is to process their emails as quickly as possible, without extra security checks. However, in case their organization sends to them simulated phishing messages in order to increase their security awareness, they may become disgruntled and unmotivated, or start blaming themselves for inability to make a correct decision in an ambiguous situation under time pressure.

Although our study led us to hypothesize about negative consequences of the human-centered anti-phishing defenses, we do not have enough evidence to support these hypotheses. Thus, one of the most important directions for future research is development of study designs and measurement procedures for assessing not only effectiveness of anti-phishing measures, but also their impact on the work and life environment of people, and on their psychological well-being.

3.3 Inconsistent Deception and Attribution

Matt Bishop (University of California – Davis, US)

License © Creative Commons BY 3.0 Unported license
© Matt Bishop

Deception is an ages-old tactic for confusing an adversary. In computer science, deception presents a “fiction”, or false reality, to the adversary. The adversary will then act and react based on this false image of the system, and the defenders can have the fiction respond in ways that will cause the attacker to reveal information and methods about the goals and attack techniques. This requires time and resources as well as planning for the attack and developing the fiction.

If the goal is to prevent the attacker from obtaining information, then the defenders must ensure the attackers do not know whether they have succeeded. For this, the consistency of the common fictions is unnecessary. Inconsistent deception confuses the adversary so they do not know what is true; they may know they are being deceived (and probably will), but so long as they cannot determine what is accurate, they cannot know when they have succeeded in finding or altering the information.

Attribution is a key part of defense, because the defenders want to know who or what organization(s) are behind the attack; similarly, the attackers will want to hide that information, possibly using deception to trick the defenders into misattributing the attack. An interesting and relevant question is how and when attribution should be provided, and the effects of different types and levels of assurance of that attribution co-existing on a system or network (such as the Internet).

The research questions I have that are relevant to this workshop are:

- Inconsistent deception is based on the theory that it will confuse an adversary, to the point that the adversary will go away. How would one validate or refute this theory?
- Could one frame inconsistent deception in such a way it seems like the system is flaky rather than the adversary being deliberately deceived?
- Under what conditions do the different types of attribution meet the needs of the involved (and intermediate) entities?
- How would one tie attribution to particular roles, and manage this connection, in a network like the Internet?

3.4 Research in Social Engineering


Jan-Willem Bullée (University of Twente, NL)

License  Creative Commons BY 3.0 Unported license
© Jan-Willem Bullée

I am Jan-Willem Bullee, and I am a Postdoctoral researcher at Linköping University in Sweden and a visiting researcher at Erasmus University Rotterdam in The Netherlands. In this capacity, I work closely with Prof Jeff Yan (LIU) and Dr Sophie van der Zee (EUR). During my doctoral research, I investigated social engineering (a form of cybercrime) in an organisational setting. I was particularly interested in the factors that explain and reduce victimisation of social engineering attacks. I explored three types of social engineering (i.e. face-to-face, telephone and email) in field experiments. Furthermore, I made a meta-analysis on social engineering interventions and a systematic review of the success of phishing emails. I also presented research ideas related to obtaining more insight into email phishing. For example: How can boosters be used to reduce the decay effect of an intervention; and what is the role of culture on the success of a phishing email?

3.5 The Federal Trade Commission


Joe Calandrino (Federal Trade Commission – Washington, US)

License  Creative Commons BY 3.0 Unported license
© Joe Calandrino

The Federal Trade Commission is the US government's primary consumer protection agency. The laws that the agency enforces include ones prohibiting deceptive practices in or affecting commerce. The FTC's Office of Technology Research and Investigation has a number of roles, which include conducting research relevant to the agency's mission. Our research has explored topics from email authentication to targeted advertising. Through research that helps identify, understand, and prevent potential deceptive practices, Dagstuhl attendees can help us protect consumers against such practices.

3.6 MITRE's Human Behavior and Cybersecurity Research and Capabilities


Deanna Caputo (MITRE – Washington D.C., US)

License  Creative Commons BY 3.0 Unported license
© Deanna Caputo

Cybersecurity has been primarily tackled from the technological perspective in academia, government, and industry. Focusing on the human aspects without training in the behavioral sciences reduces effectiveness. Behavioral scientists uniquely bring applied subject-matter expertise in human behavior to cybersecurity challenges. MITRE, as a not-for-profit who manages federally funded research and development centers, leverages human behavior to reduce cybersecurity risk using the behavioral sciences to understand and strengthen the human firewall through its Human Behavior and Cybersecurity Capability area. We utilize operational research and consultations, as well as direct sponsors' unpublished best practices across projects and portfolios to improve government and national critical infrastructure, particularly insider threat, usable security & technology adoption, cyber risk perceptions & awareness, cyber exercises & teams. Currently, we have been tasked with creating a data-driven insider threat framework that includes psycho-social and cyber-physical characteristics that could be common, observable indicators for insider attacks. Existing frameworks ignore psycho-social characteristics or are based on poor quality data. MITRE will receive, store, structure, hand-code, aggregate, and analyze a large dataset (5-10K) of raw insider threat case investigation files shared directly from multiple organizations. The framework will include: insider attacker's actions before, during, and after an attack; individual-level factors (e.g., role, character, stressors, motivations, intent); organizational factors (organizational procedures, infrastructure elements, security elements, peer information, sector); and key flags and events that led to major decisions in the inquiry/investigation. In addition, to counter the issue of underreporting of insider risks using human sensors, MITRE has conceptualized and developed an Insider Risk Personas Methodology aimed at helping government and critical industry infrastructure to operationalize insider risk in a manner that is relevant, tangible, time-practical and expandable to supervisors/HR. The outcome of the methodology is a set of evidence-based personas that are designed to help supervisors directly challenge the rationalizations that they offer for under-reporting employee risks, increase supervisor confidence and good judgments of employee risk, and increase employee risk reporting in terms of both frequency and quality. We are currently developing and will test and evaluate a set of insider risk personas specifically for the financial critical infrastructure sector. Other problem areas for multi-disciplinary (not interdisciplinary) collaboration between the behavioral and cybersecurity sciences include: imposing costs on cyber threat actors, changing cyber adversary behavior, measuring cybersecurity awareness programs, and the impact of cyberattack response/recovery on public perception/trust.

3.7 Towards Cognitive Security


Claude Castelluccia (INRIA – Grenoble, FR)

License  Creative Commons BY 3.0 Unported license
© Claude Castelluccia

My talk was about Cognitive Security. We tend to think of cyber-attacks, or cybersecurity in general, as network intrusions, malware, Denial Of Service (DOS) attacks or other exploits that compromise physical infrastructures. However recent events, such as the Russian interference attacks on the US election, have shown that humans are increasingly becoming the targets of attacks. Instead of attacking infrastructures, adversaries are using information and existing services, such as social networks, to manipulate people. Adversaries attack humans via weaponized information. Information disorder has evolved from a nuisance into high-stakes information war. It is urgent to secure our “cognitive infrastructure”. My talk discussed the foundations of the field of cognitive security. I presented a systematic analysis framework to help scientists and policy makers to tackle the topic. More specifically, the proposed framework combines the IP (Information Processing) model, used in cognitive psychology, together with the CIA (Confidentiality, Integrity, Availability) triad, used in information security, to conceptualize the field of cognitive security. Although this approach might seem simplistic and should not be taken literally, we believe it provides a useful framework to start building the foundations of cognitive security.

3.8 Measuring Online Radicalisation

Yi Ting Chua (University of Cambridge, GB)

License  Creative Commons BY 3.0 Unported license
© Yi Ting Chua

My presentation focused on the topic of online radicalisation. Using repeated measures analysis of variance (RM-ANOVA) and social network analysis, the study found changes in expressed ideological beliefs both at the forum and individual level. Specifically, differential reinforcement and differential association were the strongest predictors towards changes in expressed far-right ideological beliefs which include beliefs such as xenophobic, anti-semantic and anti-taxation.

3.9 The Neurobiology of Financial Abuse

Natalie Ebner (University of Florida – Gainesville, US)

License  Creative Commons BY 3.0 Unported license
© Natalie Ebner

Financial abuse is one of the most common forms of elder mistreatment, with devastating consequences. A rapidly aging population, combined with changes in decision making, render fraud targeting older adults a public-health concern. Technological advances open novel avenues for fraud. Older adults increasingly navigate the Internet and are at increased risk of becoming victims of cyber social-engineering attacks, such as phishing emails, which lure users into visiting webpages that procure personal information or into clicking on malicious

links. We adopted an ecologically valid approach to uncover age-related vulnerabilities in trust-related decision making. Study 1 recorded browsing activity over 3 weeks, during which young and older participants, unbeknownst to them, received simulated phishing emails. Close to half of the users were susceptible to phishing, with older women most vulnerable. There was a discrepancy, particularly among older users, between self-reported susceptibility awareness and behavior. Examining specific risk profiles, higher susceptibility was associated with lower memory and positive affect among the oldest users. In a complementary study, we contrasted brain structure and function in older adults who were victims of fraud with older adults who had avoided an attempted fraud. The exploited group showed cortical thinning in anterior insula and reduced functional connectivity within default and salience networks, while increased between-network connectivity. Thus, alterations in brain regions implicated in trust-related decision making may signal heightened fraud risk in older adults. Our data advance understanding of brain and behavioral processes underlying age-related vulnerabilities to fraud online and in-person. Determination of cognitive, socio-affective, and neurobiological risk profiles is crucial to develop prevention against victimization in aging, which can have dramatic consequences for the individual and society.

3.10 Research in Online Fraud


Matthew Edwards (University of Bristol, GB)

License  Creative Commons BY 3.0 Unported license
© Matthew Edwards

In this brief introductory presentation, I discussed elements of my research background which were related to the topic of this Dagstuhl seminar: my work on persuasion in 419 email scam exchanges, detecting online dating fraud profiles and ongoing work investigating cybercriminal fora. While I am a computer scientist, my work has been carried out in close collaboration with psychologists, and psychology informs a lot of my research. In the work on 419 scam exchanges, we have been looking at the traces of persuasion principles we can observe by looking at the text of scambaiter and victim interactions with scammers – some of which are extraordinarily long-lived. Our work on dating fraud profiles built upon suggestions that users with more romantic naiveté were more likely to become victims, building automatic classifiers that distinguish between the profiles of scammers and real dating site users. In my ongoing work, I am looking at evidence about the characteristics, historic impact, and careers of cybercriminals in underground forums.

3.11 The Sociology of Phishing

Freya Gassmann (Universität des Saarlandes, DE)

License  Creative Commons BY 3.0 Unported license
© Freya Gassmann

In general, my topics are university research, IT-Security, sport sociology and methods in social science. As a sociologist I am interested in the social science part of IT-security and quantitative data collection and analysis methods. In the last years I worked together with Zina on some projects. The last two papers were about phishing and we tried to figure out, why people click on a link in an email or Facebook message. In a field experiment over 1200

university students received an email or a Facebook message with a link to (non-existing) party pictures from a non-existing person. In a questionnaire there were asked about their clicking behavior. The most frequently reported reason for clicking was curiosity followed by the explanations that the message fit to the circumstances of the participants.


I am interested in the following questions: Why do people act risky (data protection and phishing). Are they careless or do they don't understand the importance of data and data protection? If this would be the case: Do we need better and more education for children, young adults and employees?

References

- 1 Gassmann, F., Benenson, Z., & Landwirth, R. *Kommunikation als Gefahr: Nutzerreaktion auf Nachrichten mit verdächtigen Links per E-Mail und Facebook* Österreichische Zeitschrift für Soziologie, 44(S1), 135–155, 2019 <https://doi.org/https://doi.org/10.1007/s11614-019-00351-6>
- 2 Benenson, Z., Gassmann, F., & Landwirth, R. *Unpacking Spear Phishing Susceptibility*. In M. Brenner, K. Rohloff, J. Bonneau, A. Miller, P. Y. A. Ryan, V. Teague, . . . R. Landwirth (Eds.), *Financial Cryptography and Data Security. FC 2017. Lecture Notes in Computer Science*, vol 10323 (pp. 610–627), 2017
- 3 Gassmann, F., Beck, J., Gourmelon, N. & Benenson, Z. What is more valuable: Confidentiality or availability of data? Work in progress on an online experiment using willingness to pay (WTP) in a ransomware scenario to examine users' valuation of their data. Poster at the conference of the "Akademie für Soziologie". *Digitalsocieties2019*, September 25-27, Konstanz, 2019

3.12 Caught in the Crossfire / The language of aggression, violence, and cybercrime

Alice Hutchings (University of Cambridge, GB)

License  Creative Commons BY 3.0 Unported license
© Alice Hutchings

eWhoring is a term used by offenders to refer to social engineering techniques where they imitate partners in virtual sexual encounters. Victims are asked for money in exchange for pictures, videos, or sexual-related conversations. The harms associated with eWhoring, which involves fraud by misrepresentation, include the exploitation of those being impersonated, usually young women. Some of the images being distributed are indecent images of children, or material leaked as 'revenge porn'. My previous research provides a crime script analysis of eWhoring, identifying the steps involved, the types of actors, and points for intervention. However, one of the concerns about the intervention approaches developed was the impact on those caught in the crossfire. It is important to consider the impact of crime prevention initiatives on the law abiding majority. In some cases, this may cause additional nuisance, such as the time and effort required for account verification. In other cases, it may have particularly adverse impacts on those already marginalised. In the context of eWhoring, this includes those involved in legitimate sex work, particularly if their images are stolen and used fraudulently.

In relation to aggression, my colleagues and I found that the language used on Hackforums was less aggressive than Wikipedia page edit comments. This is perhaps due to its relatively homogenous population. Targets for harassment are likely to be located off, rather than on,

the forum. However, the language used by the computer security is also interesting. Despite cybercrime being relatively non-physical in nature, the language used to describe it is often borrowed from the areas of aggression and violence. For example, we refer to incidents as ‘attacks’, and targets being ‘hit’. ‘Hacking’ has relatively sinister connotations, as does ‘defacing’. There are further examples: ‘brute force’, ‘penetration testing’, ‘smashing the stack’, ‘bomb’ (e.g. logic, fork, zip), ‘Heartbleed’, ‘Rowhammer’, ‘Shellshock’, ‘Bashbug’, and even cyberwarfare. Does this represent something about the way we perceive cybercrime? Does it relate to the way it is represented, is it framed in such a way to be considered newsworthy? Or perhaps it reflects the relative masculinity of the computer security industry?

3.13 Psychological aspects of Cybercrime

Marianne Junger (University of Twente, NL)

License © Creative Commons BY 3.0 Unported license
© Marianne Junger


First, in my presentation I have presented slides based on research on the origins of aggression in humans [1]. I stated that aggression is an innate drive in humans. Accordingly, it is ever-present behavioral option, starting at birth. Therefore, aggression has to be unlearned in childhood and this needs to be done before age 8. This unlearning process is done through a socialization process by parents and teachers. The result is that children are taught self-control. After age 8, behavioral tendencies remain relatively stable over life [2, 3, 4]. The level of self-control that has been reached has many implications. First, humans differ on self-control, not everyone has been socialized equally well. Probably genetic differences may make some children a little harder to socialize. Also, with low-self-control, humans are prone to commit all sorts of deviant behaviors, that is, all sorts of crimes and all types of risky and unhealthy behaviors. Second, I mentioned that humans have ‘truth bias’ [5]. This bias facilitates crime victimization.

References

- 1 Tremblay, R.E., Developmental origins of disruptive behaviour problems: the ‘original sin’ hypothesis, epigenetics and their consequences for prevention. *Journal of Child Psychology and Psychiatry*, 2010. 51(4): p. 341–367.
- 2 Olweus, D., Stability of aggressive reaction patterns in males: A review. *Psychological Bulletin*, 1979. 86(4): p. 852-875.
- 3 Piquero, A.R., et al., Stability in aggression revisited. *Aggression and Violent Behavior*, 2012. 17(4): p. 365-372.
- 4 Heckman, J.J., Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 2006. 312(5782): p. 1900 – 1902.
- 5 Burgoon, J.K. and T.R. Levine, Advances in deception detection. *New directions in interpersonal communication research*, 2010: p. 201-220.

3.14 Research in Security Risk Management

Katsiaryna Labunets (TU Delft, NL)

License  Creative Commons BY 3.0 Unported license
© Katsiaryna Labunets

My background is in cyber risk management and empirical research. In my PhD thesis, I conducted an empirical comparison of security risk assessment methods and investigated the criteria behind methods' success. However, cyber risk management based just on technical solutions cannot provide 100% security to organisations. Therefore, in the past years, my research focus is on how combined security measures can effectively manage human-related threats. My future research interests include security behaviour definition from organisation management and employees perspective and how actual security behaviour can be measured and explained.

In my talk at Dagstuhl, I proposed a few ideas for the workshop:

- Use a honeypot network to catch, study and suppress cyberbullies;
- Apply a serious gaming approach to train adults about cyberbullying and how to deal with this;
- Develop a catalogue of social/human-specific cyber threats and related countermeasures that can become a part of an information security standard and used by existing cyber risk assessment methods.

3.15 Research in Phishing


Elmer Lastdrager (SIDN Labs – Arnheim, NL)

License  Creative Commons BY 3.0 Unported license
© Elmer Lastdrager

In this introduction talk, I discussed my PhD research on phishing. Specifically, I discussed studies on thinking out loud, teaching children how to recognise phishing emails and websites, and a brief overview of analysing 700.000 phishing emails. After that, I discussed my research interests in Internet of Things (IoT), which cover both technical solutions (e.g., analysing network traffic) and user-oriented solutions (e.g., improving user cyber hygiene). The last part of the introduction talk was a list of ideas for future research.

3.16 Phishing Susceptibility as a Function of Age, Gender, Weapon of Influence, and Life Domain

Daniela Oliveira (University of Florida – Gainesville, US)

License  Creative Commons BY 3.0 Unported license
© Daniela Oliveira

Phishing is key in many cyber attacks. Successful emails employ psychological weapons of influence and relevant life domains. I discussed my research on phishing susceptibility as a function of Internet user age (old vs young), weapon of influence, and life domain. I presented results from a 21-day study conducted with 158 participants (younger and older Internet users). Data collection took place at the participants' homes to increase ecological validity.

Our results show that older women were the most vulnerable group to phishing attacks. While younger adults were most susceptible to scarcity, older adults were most susceptible to reciprocation. Further, there was a discrepancy, particularly among older users, between self-reported susceptibility awareness and their behavior during the intervention. Our results show the need for demographic personalization for warnings, training and educational tools in targeting the specifics of the older adult population

3.17 Cyber Deception and Cyber Aggression

Simon Parkin (University College London, GB)

License  Creative Commons BY 3.0 Unported license
© Simon Parkin

In this talk I discuss two domains of research. Regarding cyber deception, I focus on cyber-enabled fraud and its impact on smaller charities and businesses; organisations such as these may not have sophisticated cybersecurity capabilities to defend from cyber-enabled fraud. I speculate that we may be able to develop capabilities to support these kinds of organisations to assess trustworthiness, and to assess online indicators of trust (and mistrust), which is critical given the importance of trust to how charities and businesses operate online and in electronic communications. Regarding cyber aggression, I highlight challenges in mitigating technology-enabled domestic abuse and violence ('tech-abuse'). Consumer devices may be used to coerce, monitor, or control another person in a shared environment, potentially using standard device features. The capabilities of emerging Internet-of-Things (IoT) devices may have implications for those impacted by interpersonal abuse, as devices such as 'smart' locks and thermostats may be manipulated. This raises questions as to where technology can, and cannot, address related harms of abuse, but also whether there are opportunities for technology to better support those who are able to leave an abusive situation.

3.18 Get to know your geek: towards a sociological understanding of incentives to develop privacy-friendly free and open source software

Stefan Schiffner (University of Luxembourg, LU)

License  Creative Commons BY 3.0 Unported license
© Stefan Schiffner

Overall, we observe a political will that resulted in legislation that mandates developers to provide privacy friendly and secure software. Moreover, when directly asked, software developers do claim that they want to provide secure products. However, privacy incidents are still on the rise and often criminals abuse insecure implementations for their gain. We road map research for a better understanding of software developers motivations and how to create more effective legal incentives for more secure software. For now, we sketched game theoretical model. In a next step we will obtain data through qualitative and quantitative research in FOSS (free and open source software) developer community. This collected data will be used to develop an objective function for a social game. We will use these games to further analyze the current situation in the field of FOSS wrt privacy features. Lastly we will use our findings to propose changes in policy and best practice.

3.19 Characterizing Disturbing and Reactionary Content in Youtube


Michael Sirivianos (Cyprus University of Technology – Lemesos, CY)

License  Creative Commons BY 3.0 Unported license
© Michael Sirivianos

Social networking services have been affected by disinformation, manipulation, and inappropriate content. One of the most popular OSN platforms is Youtube, where a large number of the most-subscribed channels target children of a very young age. While much of this content is age-appropriate, there is also an alarming amount of inappropriate material available. However, Youtube’s algorithmic recommendation engine raises many questions related to the “rabbit hole effect”, “echo chambers”, and other issues. Furthermore, extremists participate extensively in social networks, expressing their aggressive contents and beliefs. They have an outsized impact in communities, campaigns, and political events. For example, Incels have emerged as one of the most influential extremist communities. They define themselves as unable to find a romantic or sexual partner despite desiring one. While being manifestly sexist, their ideology combines various racist and reactionary elements. They express their hate through forums and mainly on videos especially on Youtube. Sovereign citizens are another group of extremists. Any law of the state is rejected by them, they protest taxation, and in the most extreme case, they act violently, usually against government officials. Alarmingly, pedophiles also form communities around YouTube videos. As these problems persist and grow in size, states are called upon to regulate content moderation in social networks.

3.20 Characterization, Detection and Mitigation of Antisocial Behaviour

Ivan Srba (STU – Bratislava, SK)

License  Creative Commons BY 3.0 Unported license
© Ivan Srba

Growing negative consequences of online antisocial behavior in social media (e.g., fake news, rumours, hating, trolling) have recently elicited many research efforts, aimed at characterization, detection as well as mitigating of this undesired behavior. In our projects REBELION (<https://rebellion.fiit.stuba.sk/>) and MISDEED (<https://misdeed.fiit.stuba.sk/>), we aim to solve a part of open problems related to online antisocial behavior, which persist despite a large body of already existing research. In particular, the main research challenges, that we are addressing, are: 1) a large amount of unlabeled and dynamic data (the existing datasets are static and either too small or labelled by very simplified heuristics), 2) a more extensive utilization of data about content, users and context (the existing methods do not take advantage of the whole spectrum of available data, such as multiple modalities, data from multiple platforms), and 3) a proposal of new mitigation approaches (there is a need for early detection and more extensive involvement of users). In order to obtain suitable data needed to address these research challenges, we proposed and developed a unique platform for monitoring antisocial behavior called Monant [1]. It consists of several modules for web monitoring, integration of various AI methods, platform management as well as a module for providing results to end users (public and experts). In order to evaluate this platform, we conducted a case study in which we monitored 29 unreliable medical news sites and blogs. We obtained about 58 thousand news articles, which we mapped to 131


cancer “treatments” (adapted from the list provided in [2]) which have not been proven to actually cure the patient. A case study revealed us how many articles share the most frequent misinformative treatments and the time evolution of their spreading. In our future work, we plan to work on additional development of Monant platform, gathering a more extensive dataset of medical misinformation, labelling the dataset by a claim presence and stance detection, developing detection methods for various types of antisocial behavior, which will take advantage of feature-rich data provide by the dataset, and finally we will investigate new mitigation strategies, which will be deployed in Monant end-user applications.

References

- 1 Ivan Srba, Robert Moro, Jakub Simko, Jakub Sevcech, Daniela Chuda, Pavol Navrat, Maria Bielikova. Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. Workshop on Reducing Online Misinformation Exposure – ROME 2019. July 25, 2019, Paris, France.
- 2 Amira Ghenai, Yelena Mejova. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW), 2018.

3.21 Measuring and Modeling the Online Information Ecosystem

Gianluca Stringhini (Boston University, US)

License  Creative Commons BY 3.0 Unported license
© Gianluca Stringhini

The online information ecosystem is complex, with users using multiple online services at the same time, each with its own characteristics. To properly study how malicious activity unfolds on the Web, we need tools that enable us to collect data from these services at scale and enable us to get a comprehensive view of the activity happening on them. To this end, together with my group I developed a number of techniques that enable us to collect data about malicious online activities. Such techniques include developing account honeypots (e.g., on Gmail) and leaking credentials to them so that we can observe how criminals interact with them [1], setting up crawlers for online services, and leveraging social network APIs to collect data in real time [2]. I have then used this data to better understand several types of malicious activity, from cyberbullying [3] to disinformation [4]. Studying these phenomena presents a number of challenges. First, human driven malicious activity (for example cyberbullying) tends to be more nuanced and context dependent than automated one (for example spam), and therefore develop systems to automatically detect it is more challenging. To address this challenge, in my research I apply a mixed method approach in which human annotators label content that is later processed by machine learning techniques [5]. Second, online information is not only conveyed through text, but also through images and videos. To take this into account, in my research I apply image processing techniques to understand how images are used to spread hateful content online [6, 7]. Finally, online services do not operate in a vacuum but information from one service is shared on and can influence other services. To address this challenge, in my research I develop methods to keep track of influence between different online services (e.g., Hawkes Processes) [4, 6].

References

- 1 Onaolapo, J. Mariconti, E. Stringhini, G., What Happens after you are Pwnd: Understanding the Use of Leaked Webmail Credentials in the Wild, ACM SIGCOMM Internet Measurement Conference, 2016.
- 2 Hine, G., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., Blackburn, J., Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web, AAAI International Conference on Web and Social Media, 2017.
- 3 Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., Mean Birds: Detecting Aggression and Bullying on Twitter, ACM Web Science Conference, 2017.
- 4 Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., Blackburn, J., The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources, ACM SIGCOMM Internet Measurement Conference, 2017.
- 5 Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N., Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior, AAAI International Conference on Web and Social Media, 2018.
- 6 Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., Blackburn, J., G. Suarez-Tangil, On the Origins of Memes by Means of Fringe Web Communities, ACM SIGCOMM Internet Measurement Conference, 2018.
- 7 Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Luque Serrano, J., Stringhini, G., “You Know What to Do”: Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks , ACM Conference on Computer-Supported Cooperative Work and Social Computing, 2019.

3.22 Disinformation as a Political Game

Gareth Tyson (Queen Mary University of London, GB)

License  Creative Commons BY 3.0 Unported license
© Gareth Tyson

The presentation explored the role of political influence and decision making within the regulation of social media companies. We defined politics as the activities associated with the governance of a country or area, especially the debate between parties having power. This provided an underpinning for exploring how disinformation, and its subsequent regulation, can be best modelled as a political game. In most cases, we found that social media companies shy away from public power, distancing themselves from the responsibilities that it entails. In sum, this leads to a lack of accountability and problems in defining liability for harms derived from disinformation. The presentation concluded with two open-ended questions: 1) who should be given the power decide what misinformation is? and 2) what methods to enforce those decisions should be given?

3.23 Language-based deception detection

Sophie van Der Zee (Erasmus University – Rotterdam, NL)

License  Creative Commons BY 3.0 Unported license
© Sophie van Der Zee

Language use is affected by deception. For example, when lying, people distance themselves more by using more third person pronouns. Usually, this type of research is done on single statements made by many individuals. This time, we analyzed many statements made by one single individual: The US President. Thanks to the fact-checking efforts from the Washington Post, for the first time in history, there are enough fact-checked incorrect statements made by one individual to create a personalised model of deception. We collected 3 months of tweets by @realDonaldTrump, and connected this datafile to the fact-checked database by the Washington Post. We compared language use with LIWC software between factually correct and incorrect tweets. If the US President was aware of the incorrectness of his statements at the moment of sending, one would expect language difference between correct and incorrect tweets in line with the deception literature (deception hypothesis). If the US President was unaware of the incorrectness of his tweets at the moment of sending, little language differences between factually correct and incorrect tweets are expected (misinformation hypothesis). Results showed that almost half of the LIWC word categories differed between his correct and incorrect tweets, suggesting the US President is often aware that his factually correct and incorrect messages are different at the moment of sending, supporting the deception hypothesis. Next, we estimated a logit model to test how well we could predict whether a tweet was factually correct or incorrect based solely on word use. We collected a second dataset, again comprised of three months of tweets by the US President. Both within- and out-of-sample testing results led to a prediction overall accuracy of 73%. In other words, we can correctly predict for 3 out of 4 tweets by the current US President whether it is factually correct or incorrect solely based on word use.

3.24 Applying Routine Activity Theory to Cybervictimization: A Theoretical and Empirical Approach

Sebastian Wachs (Universität Potsdam, DE)

License  Creative Commons BY 3.0 Unported license
© Sebastian Wachs

In my presentation, I proposed the Routine Activity Theory (RAT) as a theoretical framework for cybervictimization among adolescents. RAT has been developed by Cohen and Felson and aims to describe conditions that are favorable for crime [1]. According to the RAT, the following three essential elements must converge for a crime to occur [1]: A likely offender, absence of capable guardians, and a suitable target. I also presented briefly current analyses in which I tested the RAT empirically. In this study, I analyzed whether parental mediation of internet use (absence of capable guardians) is directly as well as indirectly via online disclosure (suitable target) associated with cybergrooming victimization. There sample consisted of self-reports from 5,938 adolescents from six countries ranging in age from 12 to 18 (M=14.77, SD=1.60). Applying mediation test using the structural equation modeling framework I found that parental mediation, online disclosure and cybergrooming victimization are directly associated. While instructive parental mediation is negatively related with online disclosure


and cybergrooming victimization, restrictive mediation is positively related to both. In addition, online disclosure partially mediates the relationship between parental mediation and cybergrooming victimization. While this analysis confirms the general usefulness of applying the RAT to cybergrooming the findings also highlight the need to educate parents to use certain strategies of mediation and inform adolescents to avoid disclosing online too much private information in the course of prevention programs.

References

- 1 Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, (44), 588–608.

3.25 Research in Evidence-based Security

Victoria Wang (University of Portsmouth, GB)

License  Creative Commons BY 3.0 Unported license
© Victoria Wang

I particularly enjoy applying scientific rigour and academic analysis to real world situations to obtain evidence-based solutions. My current research ranges over cyber/information security, surveillance studies, social theory, technological developments and online research methods. My latest research projects involve: i) data release and its related issues of trust, privacy and security [1, 2]; ii) security threats and management measures in organisations [10]; iii) formal methods for monitoring, data collection and interventions [6]; iv) a general formal theory of digital identity and surveillance [5]; v) developing new techno-social theories such as ‘Phatic Technologies’ as conceptual tools to understand cyberspace and its security issues [6, 7]; vi) cybercrime and threats in various countries, e.g., Nigeria, and various networks, e.g., the Darknet [3]; and vii) cyberbullying [4, 8]. My future research interests include – developing cyber security solutions for critical infrastructure, and developing my Phatic Technology Theory for applications in marginalised urban societies.

References

- 1 Wang, V. & Shepherd, D., Exploring the extent of openness of open government data – A critique of open government datasets in the UK (accepted – *Government information quarterly*)
- 2 Wang, V., Shepherd, D. & Button, M., The Challenges of Opening Government Data in the UK – A View from the Bottom. *Information Polity*, 24(1), 2019: 59-74.
- 3 Mirea, M., Wang, V. & Jung, J., The not so Dark Side of the Darknet – A Qualitative Study. *Security Journal*, 32(2), 2019: 102-118.
- 4 Edwards, S. & Wang, V., There are Two Sides to Every Story – Young People’s Perspective of Relationship Issue, *Journal of Youth Studies*, 21(6), 2018: 717-732.
- 5 Wang, V. & Tucker, J.V., Surveillance and Identity: Conceptual Framework and Formal Model. *Journal of Cybersecurity*, *Cybersecurity*, 3(3), 2017: 145-158.
- 6 Johnson, K., Tucker, J.V. & Wang, V., Theorising Monitoring: Algebraic Models of Web Monitoring in Organisations. In P. James, M. Roggenbach (eds), *Recent Trends in Algebraic Development Techniques*, 23rd International Workshop, WADT 2016, Revised Selected Papers, *Lecture Notes in Computer Science (LNCS)*, Springer, 10644, 2017: 13-35.
- 7 Wang, V. & Tucker, J.V., Phatic Systems in Digital Society, *Technology in Society*, 46, 2016: 140-148.

- 8 Wang, V. & Edwards, S., Strangers are Friends I haven't Met Yet: A Positive Approach to Young People's Use of Social Media, *Journal of Youth Studies*, 19(9), 2016: 1204-1219.
- 9 Wang, V., Tucker, J.V. & Haines, K., Phatic Technologies in Modern Society, *Technology in Society*, 34 (1), 2012: 84-93.
- 10 Cyber Security Breaches Survey (2016-2019), Commissioned by HM Government (Department for Business, Innovation & Skills), Ipsos MORI.

3.26 Deception and deterrence

Jeff Yan (*Linköping University, SE*)

License © Creative Commons BY 3.0 Unported license
© Jeff Yan

What I have looked into include deception, social engineering, cybercrime and usable security, and we're interested in both technical and sociotechnical aspects. The project on "Deterrence of deception in sociotechnical systems", funded by EPSRC, enabled some exciting research and interaction with brilliant minds including Ross Anderson, Nick Humphrey, Aldert Vrij, Jeff Hancock, Jussi Palomäki and Sophie van der Zee. One of the innovations was a naturalistic behavioural study of Machiavellian individuals on strategic deception. Inspired by Oxford research on the Sicilian mafia, my recent cybercrime study examined the phenomenon of 'scam villages' in China from an economics perspective. My earlier research studied cheating in online games. Research questions which I am curious about and would like to get inspiration for in this week are abundant, for example:

- Deception deterrence: which context, and how?
- Cheat & show-off, or cheat but hide? Is Bernard Madoff the exception, or the norm? Any theory, in psychology, criminology or whatever, explaining either way?
- What research will both CS and social scientists like?
- What is the next big question?

4 Working groups

4.1 Theme 1: Attacker Modeling Group

Abhishta Abhishta (University of Twente, NL), Zinaida Benenson (Universität Erlangen-Nürnberg, DE), Matt Bishop (University of California – Davis, US), Joe Calandrino (Federal Trade Commission – Washington, US), Natalie Ebner (University of Florida – Gainesville, US), Manuel Egele (Boston University, US), William Robertson (Northeastern University – Boston, US), Victoria Wang (University of Portsmouth, GB), and Savvas Zannettou (Cyprus University of Technology – Lemesos, CY)

License © Creative Commons BY 3.0 Unported license
© Abhishta Abhishta, Zinaida Benenson, Matt Bishop, Joe Calandrino, Natalie Ebner, Manuel Egele, William Robertson, Victoria Wang, and Savvas Zannettou

The desired outcome of the group was to determine how to develop one or more probabilistic models that will predict what attackers will do next, or augment the defenses to slow down the attacker, or speed up the defense to handle the attacker better.

The group decided to focus on organizations, because they have some sort of a management plan, giving them coherence and one or more general purposes; they also have different, often complex, technological structures. Although the majority of group members were technical, the group included a criminologist and a psychologist. The group realized that any model developed had to include non-technical factors.

There are a number of ways to develop such probabilistic models. The first is to design a model based on expertise and experience, and then use data to test its accuracy. The alternate approach is to reverse this: gather real world data and develop a model based on that. In this latter case, the model would then be tested against out of sample data. The data will consist of data from attacks, data from defenses is important here, because that data provides both contextual information about the environment, i.e., the organisations involved, and the attack itself, as well as the policies and procedures the defenders use to contain (e.g., minimising its potential damage) or thwart the attack. The procedures here will be those that are used in practice, not simply the ones written in guidelines that the security management (both technical and human) personnel and users are supposed to follow.

This leads to the first step: obtaining real world data required to build such a model. It is unclear at this point what attributes the data must have, and indeed what the data itself must consist of, so an appropriate approach is to see what data is available now, what it consists of and what attributes it has. As the model is developed and refined, aspects of the data and attributes that are missing and yet are necessary for the model to predict effectively will become clear. Also, techniques for obtaining the data are essential, because while much data has been gathered, very little of it is widely available, or indeed available except under the most stringent conditions. In short, even if such data is available, getting access to the data is yet another difficult step. For example, organisations might not want to admit that they have been victimised by cyber attackers. Even if they openly admit to victimisation, they might not be willing to share their log files and other internal documents recording the attacks with researchers. In fact, based on our previous experience, this is rather common, especially within the financial and insurance industries, wherein peer competitions are intense. Thus, an open question is how to relax these constraints while providing the guarantees that the possessors of the data will require in order to share it. This ties into the ethics of gathering data, which vary among legal jurisdictions and types of organizations. For example, in the United States, public institutions must comply with one set of government rules regarding protection of personally identifiable information, whereas private entities comply with a different (but overlapping) set of rules. For another example, the introduction of the GDPR (2018) in Europe might, on the one hand, mean that organisations are under more pressure to share their data; whereas on the other hand, they might become even more cautious in sharing data with researchers.

The data will come from several sources. Technical data will come from places such as logs, network traces, and network- and host-based data; it will include contextual information such as metadata, the organization where it is gathered from and that generated it (which may be different organizations), and the location of the data and its use (for example, if it is stored in a cloud, or stored in encrypted form locally or in a cloud, and whether the computations are done locally or in the cloud, and so forth). Red teaming, also known as penetration testing, will also be a valuable source of data. Less technically complete data will inform motivations, external characteristics of the attack, and other human and organizational aspects of the data. News stories will be a good source of this type of data, as well as law enforcement reports, government analyses, and court records. Relevant questions here relate to the broader picture of attacks. How do attackers advertise their wares? What

are their wares – what tools and methodologies do they use, and do they share or sell these? Further, a series of empirical work might be conducted to gather data from employees of selected organisations, via common social science research methods, such as questionnaires, interviews, and focus groups. For example, we could simply ask employees of an organisation what they think they did right to minimise any possible damage of a recently experienced cyberattack. Here, relevant questions might be: what was their first response? Did the organisation have a Chief Information Security Officer who discover the attack and respond to it very quickly?

The group noted that, in addition to conventional cybersecurity attacks, the above may apply to the dissemination of fake news (defined as news that contains information that is verifiably untrue). The basis for this belief is that Facebook, Twitter, and other social media can be considered large-scale distributed logs, and the organization these logs apply to is the society involved.

This led to a discussion of high-level considerations. Attackers may have many goals, such as getting money, embarrassing someone or some entity, obtaining control of a system (technical or non-technical, such as a political organization) to change things (such as the politics of a society, possibly by the use of fake news), and many other goals. The group agreed to focus on financial institutions to keep the work manageable. Two SWOT (Strengths, Weaknesses, Opportunities, and Threats) analyses examined both the financial institutions (specifically, banks) and the attackers. Tables 1 and 2 summarize the results of these analyses.

From this, the group began work on the model, a preliminary version of which follows. This starting point is definitely not complete. New features will be added, and some of the (existing and new) features will be empty for a given instantiation. Hence, the reader should view what follows as an outline.

The model is based upon goals, which include interrupting services, public shaming, obtaining money or denying others money, obtaining various types of power (social/cultural, political/ideological, economic, and so forth) or denying these to others, gathering information, and other possible goals. These are more detailed than the goals outlined above, and are consonant with them.

The structure of the model consists of 9 basic features:

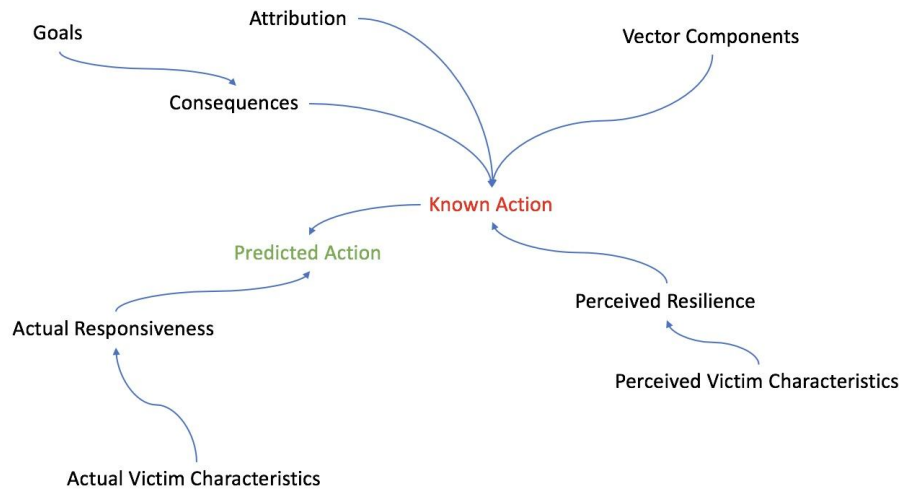
1. Goals
2. Vector components
 - a. How long does the attack take; when does it occur
 - b. Complexity of the attack (technical, non-technical, etc.)
 - c. Technological tools employed
 - d. Access (direct or indirect; social engineering, vulnerability scanning, etc.)
 - e. Communication (density, patterns, etc.)
3. Consequences (intended/unintended; who is harmed, who benefits)
4. Authorisation (authorised/unauthorised; for the latter, open or hidden)
5. Attribution (full, none, false, random)
6. Expected resilience of victim from attacker's point of view
7. Expected characteristics of victim from attacker's point of view
 - a. Location, relationships
 - b. Spread; how large is the attack surface?
 - c. Infrastructure
8. Actual responsiveness of victim
9. Actual characteristics of victim
 - a. Location, relationships
 - b. Spread; how large is the attack surface?
 - c. Infrastructure

■ **Table 1** SWOT table for financial organizations.

<p><i>Strengths</i></p> <ul style="list-style-type: none"> ■ High financial resources ■ Historically motivated to invest in security 	<p><i>Weaknesses</i></p> <ul style="list-style-type: none"> ■ Focus on financials ■ Reliance on 3rd party software ■ Legacy systems ■ Highly distributed systems
<p><i>Opportunities</i></p> <ul style="list-style-type: none"> ■ Sharing of information ■ Availability of finances ■ Substantial political capital 	<p><i>Threats</i></p> <ul style="list-style-type: none"> ■ Availability attacks on distributed systems ■ Legacy systems breaking down or compromised ■ Insider attacks (leaking of information on high profile clients) ■ Unauthorized transfers ■ Privacy issues (personal information of clients)

■ **Table 2** SWOT table for attackers of financial organizations.

<p><i>Strengths</i></p> <ul style="list-style-type: none"> ■ Force useless investment ■ Availability of cybercrime as a service 	<p><i>Weaknesses</i></p> <ul style="list-style-type: none"> ■ High resources and background information required ■ Conversion to hard cash ■ Information asymmetry
<p><i>Opportunities</i></p> <ul style="list-style-type: none"> ■ High value data ■ High value money ■ Reliance on implicit trust ■ Attack clients of the bank 	<p><i>Threats</i></p> <ul style="list-style-type: none"> ■ Getting caught (for example, when converting the electronic cash to physical cash) ■ Reputational damage to the attacker



■ **Figure 1** Relationship of the components of the structures.

Figure 1 summarizes their relationships. The model uses those structural features that drive the known action (features with paths to the red “Known Action”) to combine with the structural features that enable future actions to be predicted (features with paths to the green “Predicted Action”).

In more formal terms, a known action A_1 (which is a function of features 2-7) leads to a set of probable actions A_{21}, \dots, A_{2n} (which are a result of A_1 and feature 8). To determine the best response strategy, the net payoffs of each need to be computed. The characteristics of the victim (feature 9) drive a penalty, so the calculation must include a DB (for “DisBenefit”) component. Let $P(A_j)$ be the payoff for action A_j . From a purely rational point of view, the next action of the attacker should maximise the net payoff. Then the most likely predicted action is the one maximizing $P(A_1) + P(A_{2k}) - DB$, over k (see Figure 2). How to calculate these payoffs is left for future work. Note the assumption here is that the attacker is following some sort of rational plan; if the attack is a sequence of random actions, the underlying assumption does not hold.

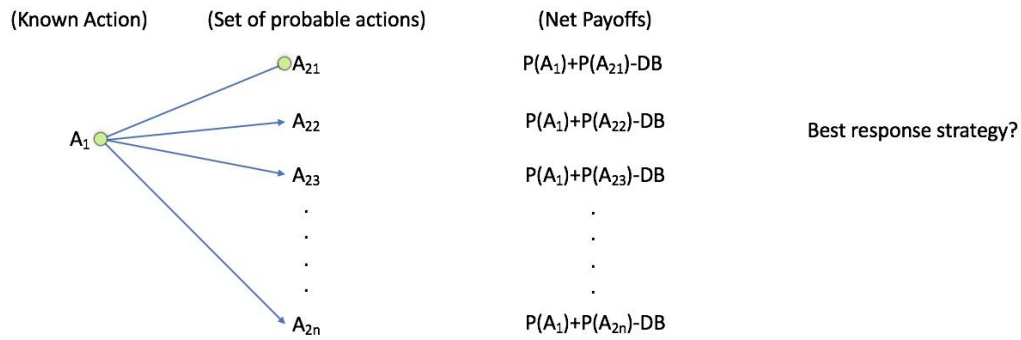
The group then used two case studies to begin validating the model. At least one member of the group worked on each of the incidents using the case studies. Tables 3, 4, and 5 summarize the application of the models to the case studies. Table 3 is the characterization of the Internet Worm of 1988; Tables 4 and 5 are the first and second steps of the SpamHaus attack.

Future work will sharpen the model and make it useful. The key opportunities for improving it are:

1. Convert the variables into measurable quantities

First, data must be found to see if the overall structure of the model works. This data can be used to determine how to measure the attributes. Undoubtedly, some will remain qualitative, and others quantitative; but the values for both types will be refined as data emerges. This will also lead to a refinement of the definitions of the variables.
2. How to obtain labeled data

Obtaining data properly labeled as attack data (as opposed to data that is unlabeled) is critical, and methods to do this must be investigated. Several possibilities were discussed,



■ **Figure 2** Graphical representation of probable actions and net payoffs.

■ **Table 3** Validating the model with the Internet Worm of 1988.

Motive	social/activism
Time	3pm-midnight
Complexity	High (considering historical context)
Technical tools	Reused software from a (suspected) re- search laboratory
Access	Direct access (to MIT public access node)
Communication	N/A (only one attacker)
Consequences	Targeted hosts on network were unusable
Authorization	Unauthorized, intended to be open
Attribution	Fully attributed
Attacker knowledge of responsiveness	None; security community nascent
Perceived victim characterization	Access to ARPANET/ARPANET/Suns and VAXen
Actual victim characterization	Access to ARPANET/ARPANET/All systems with access to ARPANET; only Suns, VAXen taken down

among them providing “data bounties” (much like “bug bounties”) and developing similar incentive structures for encouraging the sharing of data. Threat feeds may be a fertile source, as will interviews with CISOs, incident responders, and other security operations personnel. An alternative is to use the “over the fence” approach. In this approach, others take the model, instantiate it with attack data they can’t share, and give results, including problems with the model, false positives, and false negatives. This will also allow the model to be instantiated with data at different levels of coarseness.

Several open questions remain:

- Are there any higher order attacks or dimensions of attacks we are missing?
- The variables are not orthogonal – is this a problem?
- How do we handle noise in the measurements?
- How do we handle noise the attacker injects?
- How do we handle false positives/negatives? And equally if not more critical: how do we identify them?

■ **Table 4** Validating the model with the SpamHaus attack (step 1).

Motive	revenge
Time	6pm
Complexity	Low
Technical tools	Off-the-shelf tools
Access	None
Communication	IRC (after the attack)
Consequences	None
Authorization	No
Attribution	Random
Attacker knowledge of responsiveness	Low
Perceived victim characterization	SpamHaus/Not distributed/Server
Actual victim characterization	London Exchange + CloudFlare + SpamHaus/Widely distributed/CDN

The group suggested possible next steps. A workshop on defenders and attackers would provide additional insights and understanding. Such a workshop should include CISOs, security operations personnel, and others who defend systems, as well as former attackers who “came over to the light side”. Obtaining funding for this work is critical, and there was considerable discussion about what groups or agencies might fund this international collaboration. A paper on our conceptual model of cyberattacks would be a good starting point for such requests. Possible appropriate venues would be WOOT, IFIP SEC, and the economics workshop WEIS.

4.2 Theme 2: Unexpected Consequences of Countermeasures

Matthew Edwards (University of Bristol, GB), Yi Ting Chua (University of Cambridge, GB), Alice Hutchings (University of Cambridge, GB), Daniela Oliveira (University of Florida – Gainesville, US), Simon Parkin (University College London, GB), Stefan Schiffner (University of Luxembourg, LU), and Gareth Tyson (Queen Mary University of London, GB)

License © Creative Commons BY 3.0 Unported license

© Matthew Edwards, Yi Ting Chua, Alice Hutchings, Daniela Oliveira, Simon Parkin, Stefan Schiffner, and Gareth Tyson

Overview

We tackled the topic of countermeasures enacted in cybersafety and cybercrime often leading to unintended consequences and harm. This problem arises for both technical solutions (classifiers, website takedowns) and administrative solutions (staff training, public advice, policies enacted by staff). We developed a taxonomy of unintended consequences, and transformed this into a set of questions which could be asked of any countermeasure, so that potential consequences might be anticipated and mitigated.

■ **Table 5** Validating the model with the SpamHaus attack (step 2).

Motive	revenge
Time	midnight
Complexity	Low
Technical tools	Off-the-shelf tools
Access	None
Communication	IRC (after the attack)
Consequences	Drop CloudFlare from London Exchange; SpamHaus no longer reachable
Authorization	No
Attribution	Full
Attacker knowledge of responsiveness	Low
Perceived victim characterization	London Exchange + SpamHaus/Not distributed/Server
Actual victim characterization	London Exchange + CloudFlare + SpamHaus/Widely distributed/CDN

Section 1: Scenarios

The group approached the problem by first defining a set of cybersafety scenarios as motivating examples, then identifying countermeasures which may be applied to these scenarios. These countermeasures were then used as grounded prompts for consideration of unintended consequences.

1. Intimate partner abuse¹: Bob and Charlie live together. Bob is controlling and monitors Charlie's behaviour using IoT devices. This includes Charlie's smartphone. When suspecting Charlie might be visiting his friends, Bob goes onto Twitter and shares aggressive and fabricated posts.

Countermeasures & Consequences:

- Take away Charlie's tech so Bob cannot use it to harm them. Replace all of Bob's accounts with new ones.
 - Loss of personal information
 - Financial cost
 - Loss of abilities provided by tech (to stay in contact with family and friends)
- Provide training resources for Charlie so they know how to identify and prevent this abuse.
 - Bob might find this advice and become more violent
 - Bob might use this advice to become more stealthy and effective in abuse of Charlie
- Recover and reset devices – as the UK government suggests
 - Loss of personal information
 - Loss of social support structures
- In cases where intimate content is shared – contact social media company, take down material

¹ Lopez-Neira, Isabel, et al. "‘Internet of Things’: how abuse is getting smarter.", Safe –The Domestic Abuse Quarterly, (63), 22-26. Women's Aid (UK), 2019.

- Takedown mechanism might be misused to implicate innocent users
- Verifying identity might be embarrassing and/or difficult
- Streisand effect – content could become more popular
- Images might instead be shared on platforms where more harm to the victim might originate
- Legal actions – criminal prosecutions
 - Slow pace of justice system, stress
 - Risk of escalation before
- Revenge porn – facebook asks you to upload images in advance
 - Verified connection between image and your identity – future misuse
 - Normalises sharing

2. Disinformation: There is a political campaign, Charlie vs. Bob. A third party, who supports Bob, performs a concerted misinformation campaign to spread false information about Charlie. This is done predominantly via Facebook and Twitter, initiated via a network of social media bots who inject the material.

Countermeasures & Consequences:

- Remove tweets/posts
 - Backlash – spread more often in defiance
 - Takedown used as evidence of conspiracy to suppress ‘truth’
- Remove bots
 - Misclassification, irritation of innocent users
- Removing accounts
 - people move onto Gab and intensify
- Detect collusion in social graph
- Build machine learning model to identify ‘fake news’
 - Leads to complacency, reduction in skepticism
 - Misclassification
- Using fact checkers to highlight fake news
 - Costly to fact-check material
 - Complacency, trusting fact-checker for truth
- Reduce visibility of material considered to be fake news
 - Evidence of ‘suppressing truth’
 - Misclassification, innocent users don’t necessarily know what’s wrong
- Limited number of shares/forwards
 - Limit also applies to legitimate content
- Block entire service
- Promoting correct information

3. CEO Fraud: Bob finds out the name and details of the Footbook’s CEO. Bob emails one of Footbook’s employees, Charlie, asking him to pay a last minute invoice because Bob forgot. Charlie goes ahead and pays the invoice, which transfers money into an off-shore account. Charlie gets sacked.

Countermeasure:

- Change the culture of the company – CEOs can’t send random emails
 - Productivity costs, conflict
- Training

- Additional cost for the low-level employee
- Security best practice, least privilege
- Authentication required for bank transfers/third party checks on all transactions
 - Productivity costs
- Crypto check the sender
- Remove steps from email, and required in-built finance system
 - Implementation costs
- Remove domain squatting
- Automated attacks – looking for anomalous behaviour in transactions
 - Misclassification of important transactions
- Restriction of access to external sites/public email services
- If data leak (e.g. IP theft) could watermark files
 - Company use this to identify whistleblowers

4. Phishing: Bob has recently lost his job, and holds bitter resentment towards his former employer. He believes there has been a conspiracy against him, driven by mistrust of his Northern accent. He therefore formulates a phishing campaign against the HR department of his former employer. Charlie receives an email from Bob, masquerading as a notification of a company award worth £18. Charlie clicks on the link, and is asked to enter his credentials. The website, operated by Bob, is then used by him to retrieve HR data related to his dismissal. Bob was sacked because of his aggressive and inappropriate behaviour in the company toilets.

Countermeasure:

- Training & education (including phishing exercises used as training)
 - Creates a false sense of understanding the problem
 - Allows attackers to adapt to the training
 - Results in victim blaming
 - Might upset people – make them feel stupid
 - Might not help all users (e.g. ones who don't engage with training), but company might then think that the problem is solved
- Email filtering, e.g. using machine learning
 - Misclassified email goes to spam, holds up work
- Website takedown and ISP blocking of websites
 - Website takedown mechanism could be abused to take down legit sites
 - Streisand effect
 - Site might move to more resistant providers
- Website verification
 - Sense of security from verification could be misleading about behaviour
- Safe links

5. Dating Fraud: Bob is innocently swiping on Tinder. He encounters a handsome young woman, Charlie. Bob and Charlie hit it off, and instantly begin to plan their life together. Unfortunately, Charlie lives in Peru and cannot afford to travel to Dagstuhl. After a few weeks of intimate conversation, Charlie requests \$3000 to enable her to book a flight. Once the money has been transferred, Bob never hears from Charlie again.

Countermeasures:

- Get off Tinder
 - No hookups

- Verify accounts
 - Some apps force by providing link to facebook account
 - Might not want to share that information, i.e. privacy invasive
 - People with non-traditional sexual interests have them exposed
 - Might expose people to financial fraud if required to upload credit card details
- Close fraudulent accounts
 - False positives, e.g. person who is very popular
 - People may have had photos stolen from them, and used by fraudsters.
 - Countermeasures often involve collecting more data – data leaks have a greater impact
 - Might be cultural sensitivities that must be catered for, e.g. Tinder vs Grindr
- Advice, tips and prompts (targeted)
 - Annoying for users
- Training
- Waste the time of suspected scammer
 - Wastes timewaster's time/resources
 - Extended contact raises potential for more harm
 - Could provoke e.g. violence

Section 2: Taxonomy & Questions

Working from the list of consequences from countermeasures in each of these scenarios, the group categorised common types of consequence, and then reformulated the taxonomy as a number of questions which should be asked of any proposed countermeasure.

Unintended consequence taxonomy:

Additional Costs: Implementing countermeasures can pose a burden for different stakeholders involved. Training and policy exhaust employee compliance budget², restrictive security controls can hamper business productivity³, staffed reporting systems must be paid for by a social media platform.

Misuse of Countermeasure: The countermeasure itself might be misused by malicious actors to cause harm. Reporting systems can be misused to target competitors for takedown; advice for victims can be used by perpetrators to improve their misbehaviour; abusers can train against classifiers to learn how to go undetected.

False Positives: Incorrect decisions made by/as a result of the countermeasure can cause harm to innocents. Classifiers can misidentify content or users as malicious or deceptive; verification schemes can exclude people legitimately unable to verify their identity.

Displacement: The countermeasure might simply move harm to other targets. Removing extremist accounts pushes them to echo chambers where their views might be reinforced; stricter or more arcane policies may simply cause employees to circumvent policy and suffer all the blame for resulting failures.

Amplification: The countermeasure might actually cause an increase in the behaviour it intended to prevent. A plethora of fact-checkers leads to fragmentation of trust, attempts to take something down can cause it to gather more attention through controversy, harsh crackdowns can lead to reprisals in defiance.

² Beautement, Adam, M. Angela Sasse, and Mike Wonham. "The compliance budget: managing security behaviour in organisations." Proceedings of the 2008 New Security Paradigms Workshop. ACM, 2009.

³ Kirlappos, Iacovos, Simon Parkin, and M. Angela Sasse. "Learning from "Shadow Security": Why understanding non-compliance provides the basis for effective security." Proceedings of the 2014 Workshop on Usable Security (USEC). Internet Society, 2014.

Insecure Norms: The countermeasure might promote the adoption of insecure norms. Highly-trusted technology or policy can lead to a false sense of security that makes users more susceptible to deception; normalising the sharing of identifying information for verification purposes contributes to phishing success.

Disrupting other Countermeasures: A well-intentioned countermeasure could inadvertently cause problems for another – potentially more effective – countermeasure. Social media sites which remove abusive content are also removing evidence from criminal investigations; requiring users to verify their identity prevents them from using anonymity as a defence; contradictory advice on how to deal with a problem leads to confusion.

Questions

From the above categories of unintended consequence, we extract 6 questions that could be asked of any proposed (or extant) countermeasure to identify potential unintended consequences.

1. In what ways might the countermeasure burden stakeholders?
2. In what ways might the countermeasure be used in attacks?
3. In what ways might the countermeasure displace harm to others?
4. In what ways might the countermeasure amplify harm?
5. In what ways might the countermeasure create insecure norms (e.g. complacency)?
6. In what ways might incorrect classification cause harm?
7. In what ways might the countermeasure disrupt the operation of other countermeasures?

We also identify a cross-cutting concern:

8. Consider for each question which groups are more at risk of experiencing harm.

Section 3: Identifying Further Consequences

We cross-tabulated the above taxonomy with four general categories of countermeasure, to validate the location of specific unintended consequences within the taxonomy, and to make use of the taxonomy to identify new unintended harms in areas our earlier scenarios had not covered (see Table 6).

■ **Table 6** Categories of countermeasures and related unintended harms.

	Categories of Countermeasures			
	Managing content	Verification (controlling users)	Training (changing behaviours)	Takedown (infrastructure)
Displacement	Moves people to echo chambers; Fragmentation	User displacement to less protective platforms	Circumventing work policies	To abuse resistant hosting providers
Insecure Norms	Warnings; Rely on fact-checking; Normalising sharing of explicit images; Preaching to the choir; Groupthink; Non-falsifiability	Normalising sharing personally identifiable information	Makes social engineering problem routine; Desensitization; Habituation; Risk-dumping; Told wrong thing	Assume problems are removed
Additional Costs	Wiping phones; Loss of evidence; Disrupt existing connection	Annoyance / time to verify	Loss of productivity; Adds to compliance budget; Conflict; Chilling effect; Induce mistakes; Victim blaming	Criminal Justice System (slow retaliation); Legitimate sites recovery cost
Misuse	Poisoning fact-checking; Identifying whistleblowers; Sausages identify; Misuse image hashing	Privacy impacts; Misuse by; Data breach; Faking blue ticks / trust seal	Perpetrators learn from advice	Reporting competitors; Censorship
False Positive	Forcing false positives; Cold start problems – new users struggle to gain trust	Users cannot verify identity due to photo stolen	Errors as result of training	Website take-down
Amplification	Fragmentation; Streisand effect	Blue ticks on Twitter	Perpetrator sees advice and escalate	Streisand effect
Disrupting Other Countermeasures	Destroy evidence	Anonymity (e.g. Facebook and phone number)	Contradictory advice	Destroying evidence

Section 4: Directions for Research

Future research on this topic could explore a number of additional directions:

1. Do the devised questions cover enough unintended consequences that they could be used as an instrument in e.g., ethical review of security and cybersafety research proposals concerning countermeasures?
 - What are the limitations of this instrument, and can it be amended to correct for these?

2. How can the likelihood and severity of unintended harms be ascertained?
 - Can anything general be said about the likelihood and severity of the categories of unintended harms, or does this depend too much on the specific countermeasure in question?
 - How can measures of unintended consequences be gathered?
3. Why are unintended harms not already being considered?
 - Is there a facet of decision-making around countermeasures (e.g., lack of incentives) which explains why they are not considered?
 - Are they in fact not considered/seen⁴, or just too difficult to remedy?⁵
4. Are there common mitigations to unintended harms which might complement this taxonomy?
 - Can we produce guidance that allows developing countermeasures to build-in mitigations in a variety of application areas?

4.3 Theme 3: Measuring Human Behavior from Information Security and Societal Perspectives

Ivan Srba (STU – Bratislava, SK), Katsiaryna Labunets (TU Delft, NL), and Sophie van Der Zee (Erasmus University – Rotterdam, NL)

License © Creative Commons BY 3.0 Unported license

© Ivan Srba, Katsiaryna Labunets, and Sophie van Der Zee

Joint work of Ivan Srba, Katsiaryna Labunets, Sophie van der Zee, Jeff Yan, Gabriele Lenzini, Jeremy Epstein, Deanna Caputo, Jean-Willem Bullée, Claude Castelluccia

Introduction. People, organizations, and governments are increasingly using the Internet for a wide range of activities, from socializing to shopping, and working. This increased digitization has brought many benefits, but also comes with downsides. Crime is also increasingly digitized, from hate speech and cyberbullying to hacking and identity theft. Since 2016, hacking has been the most prevalent crime in the Netherlands. Specific numbers are however hard to come by. Victims of cybercrime are not reporting their victimization to the police, which leads to unreliable crime statistics. And estimations of the cost of cybercrime differ substantially between academic researchers and commercial companies offering protection, training, and insurance. In the meantime, organizations are spending much time, effort, and money on training their employees to become more resilient. However, the effectiveness of these interventions are seldom properly measured. In this working group, we aimed to identify and describe techniques to systematically measure digital behaviors relevant to the following two contexts:

1. online misbehavior and false information (e.g., fake news, rumours, hating, cyberbullying).
2. cybersecurity (e.g., phishing, ransomware).

While these online threats are commonly researched from different perspectives, we recognize a lack of well-defined and comprehensive methodological frameworks how to measure interactions between them and human behaviour – how to measure their enablers

⁴ See application of Johari Window to security activity, as in e.g., Beris, Odette, Adam Beautement, and M. Angela Sasse. “Employee rule breakers, excuse makers and security champions: Mapping the risk perceptions and emotions that drive security behaviors.” Proceedings of the 2015 New Security Paradigms Workshop. ACM, 2015.

⁵ Herley, Cormac. “More is not the answer.” IEEE Security & Privacy 12.1 (2013): 14-19.

(i.e., what makes online threats possible and effective) and influences (i.e., how online threats affect humans and their behaviour). We were particularly interested in measuring:

- reach and effect of online misbehavior and false information threats on influencing human behavior.
- security behavior that may expose individuals and organisations to cybersecurity threats.

The output of the working group are two methodological frameworks for each category of threats that consist of a list of addressed online threats and corresponding security behaviors; and identification of technical measurements that can be practically used to study such online threats and human behaviour

A framework for measuring online misbehavior and false information. As the first part of this framework, we proposed a hierarchical categorization of different types of online threats. We identified 4 main groups on online threats: deception, manipulation, aggression and mischief (we focused in more detail on the first three groups in the framework). Secondly, we identify typical victims and offenders for each online threat (as potential actors we considered individual users, communities, organizations, governments and societies). For each category of threats, we identified how we can determine:

- The reach of threats, e.g. we can measure the speed of spreading for deception and manipulation threats (such as fake news) by number of likes, retweets, shares, replies, comments, etc. per time unit.
- The effect of threats, e.g. we can measure how fake news influenced the political preference by looking at election results or changes in voting behavior.

To sum up, the framework consists of 3 main components: hierarchical categorization of different types of online threats, identification of victims and offenders and metrics to measure reach and effect of threats. In addition, we can summarize our main findings by two take-away messages: 1) We still miss a comprehensive list of definitions and categorizations for individual online threats. 2) While we can measure the reach of threats quite well (reach is well observable), the measurements of their effect cannot be determined precisely (the effect is usually hidden and influenced by a number of additional circumstances).

A framework for measuring human security behavior. A valuable input provided by Sophie van Der Zee became the basis for this work. Her initial framework consists of four components: 1) user groups, 2) possible factors that influence or can be used to influence user's security behaviour, possible 3) metrics and 4) approaches to measure user's behaviours. Based on this input, we decided to focus on possible observable security behaviours. We did a brainstorming session with group members using post-it notes and identified a list of possible security behaviours of individuals or organizational users. In this session, we came up with 26 security behaviours that we grouped in nine categories. These categories are related to browser behaviour, use of a smartphone, software, hardware, emails, passwords, document handling, laptop, and file sharing sites. In the next step, we looked into technical measurements/metrics that can be used to study the corresponding security behaviour in the wild. For example, the data describing security behaviour related to "Changing default passwords (for new accounts, routes, IoT devices)" can be collected by scanning accounts/devices based on the default password list.

As the last component of our framework, we thought about possible research study designs that can be used to investigate each security behaviour using specific technical metric. For the above example of "Changing default passwords" we proposed the following study design: "AB-test: scan for default passwords → provide awareness regarding default passwords → scan same 'population' again after a short time → compare scans."

To sum up, the framework consists of 3 main components: observable security behavior, technical measurement or metric, and suggested research study design to investigate the corresponding behavior using specific metric. This framework aims at providing researchers and practitioners with a practical and structured way of studying human security behaviours.

Conclusion. In summary, our working group drafted two methodological frameworks for researchers and practitioners who are interested in studying and measuring 1) the reach and effect of online threats on influencing human behavior and 2) actual human cybersecurity behavior.

4.4 Theme 4: Prevention, Detection, Response and Recovery

Gianluca Stringhini (Boston University, US), Freya Gassmann (Universität des Saarlandes, DE), Marianne Junger (University of Twente, NL), Elmer Lastdrager (SIDN Labs – Arnheim, NL), Michael Sirivianos (Cyprus University of Technology – Lemesos, CY), and Sebastian Wachs (Universität Potsdam, DE)

License © Creative Commons BY 3.0 Unported license
© Gianluca Stringhini, Freya Gassmann, Marianne Junger, Elmer Lastdrager, Michael Sirivianos, and Sebastian Wachs

The working group focused on Prevention, Detection, Response, and Recovery approaches with respect to cybersafety incidents. To guide the discussion, three very distinct topics were examined: (1) Cyber grooming, (2) phishing; and (3) IoT.

After the brainstorming session, the group decided to focus on a specific topic. It turned out the group was packed with expertise pertaining to research of adolescents. Therefore, the working group decided to focus on cyber grooming as the main topic for further discussions. Cyber grooming occurs when someone (often adult) befriends a child or adolescent and builds an emotional connection with future intentions of sexual abuse and/or exploitation.

The main goal of cyber grooming is to gain the trust of the child, which can for example be exploited to obtain intimate and personal data from the child (often sexual in nature, such as sexual conversations, pictures, or videos). They in turn can be used to threaten and blackmail the child for further inappropriate material or acts.

Unfortunately, adolescents rarely turn to an adult for help when they face problems online. Imposing online restrictions might be perceived as a threat to their freedom and thus induce a psychological reactance process leading to undesired behavior. In order to protect minors, we need to equip them and their guardians with appropriate tools that can tackle challenging situations and empower users to deal with threats in a thoughtful manner.

Considering all these difficulties, the group came up with the “Guardian Angel approach.” This approach entails a proper suite of cybersafety tools that let a minor create a “Guardian Avatar” which can be customized so that it feels familiar. The main goal of the guardian angel is to protect children against groomers and those who plan to abuse their trust and take advantage of them. The avatar pops up when the system detects something suspicious and advises the minor accordingly.

We mentioned a number of requirements, such as that the tool should be age appropriate and culturally appropriate. We also discussed how parents should or could be involved. Depending on the age of the minor, the system provisions for various degrees of privacy. In the first mode the avatar will be invoked only when the user initiates it. This is the least intrusive modality, which is tailored to adolescent users. In the second mode, the avatar is

automatically activated by the system in the response to intelligent detection. This modality is more appropriate for pre-adolescent children. In either case, the avatar will help the victim cope with a dangerous situation. In the case of adolescents, the system engages the minor with a series of questions, answers and advice. Parents will be notified only with the consent of the teenager. In the case of pre-adolescent children, the tool engages the parents as deemed appropriate. Moreover, the avatar-based system can become a learning environment with tutorials. Therefore, the system will also have educational value and can be introduced in classrooms and awareness workshops.

Despite the actions taken from the avatar to prevent cybergrooming, the minor may end up trusting potential attackers more than the avatar, where the minor should only trust the avatar or its parents. Therefore, special care should be taken to gain the trust of the minor by using appropriate UX design and proper settings. As a start, the parent enters the age of the minor and the system should automatically choose the right level of intervention. By analyzing the interactions with its users, the system will progressively learn how to address various types of users and situations.

We stressed that the guardian angel should, ideally, be embedded in more general policies to protect children online, such as media education at school.

Overall, the research will focus on interventions against interpersonal online aggression, ICT- based cybergrooming detection tools, seeking online help, the effectiveness of online assistants, human factors and user experience, effects of alerting parents, experiments with the monitoring of adolescent's mobile and case studies analysis. Furthermore, natural language processing, image analysis, and fact checking modules will be implemented. The evaluation will also entail three user studies which will take place within small and medium scale pilots. In particular, a study of user acceptance for the "Guardian Angel approach" will take place first. Subsequently, a group of adolescents will use the tools and cybersafety-related responses will be compared to a group that does not use the tool. In all studies, the ethnic and socio-economic background of the users will be taken into consideration.

Regarding privacy considerations, ideally the data should be processed only on the user's computer or in Web proxies at the user's residence. At the same time, feedback should be used to update the models of the project and make them more accurate. To this end, privacy-preserving federated learning approaches will be employed. Overall, the application can have various privacy preferences, ranging from 'full monitoring' for younger children, to 'on demand' for adolescents. Every action will follow the GDPR regulations and the users will be fully informed.

The above concepts will be proposed for EU-funding (probably ETN/ITN 2020). In addition, numerous stakeholders will be contacted, including the Cyprus Ministry of Education, the Cyprus Police, Adolescents' Parliament in the Netherlands, foundations (NGOs, GOs) that work with sexuality awareness for teenagers, the Dutch police, teachers, parent associations, schools, to evaluate the idea, collect feedback and to raise awareness.

Participants

- Abhishta Abhishta
University of Twente, NL
- Zinaida Benenson
Universität Erlangen-
Nürnberg, DE
- Matt Bishop
University of California –
Davis, US
- Jan-Willem Bullée
University of Twente, NL
- Joe Calandrino
Federal Trade Commission –
Washington, US
- Deanna Caputo
MITRE – Washington D.C., US
- Claude Castelluccia
INRIA – Grenoble, FR
- Yi Ting Chua
University of Cambridge, GB
- Natalie Ebner
University of Florida –
Gainesville, US
- Matthew Edwards
University of Bristol, GB
- Manuel Egele
Boston University, US
- Jeremy J. Epstein
NSF – Alexandria, US
- Freya Gassmann
Universität des Saarlandes, DE
- Alice Hutchings
University of Cambridge, GB
- Marianne Junger
University of Twente, NL
- Katsiaryna Labunets
TU Delft, NL
- Elmer Lastdrager
SIDN Labs – Arnheim, NL
- Gabriele Lenzini
University of Luxembourg, LU
- Daniela Oliveira
University of Florida –
Gainesville, US
- Simon Parkin
University College London, GB
- William Robertson
Northeastern University –
Boston, US
- Stefan Schiffner
University of Luxembourg, LU
- Michael Sirivianos
Cyprus University of Technology
– Lemesos, CY
- Ivan Srba
STU – Bratislava, SK
- Gianluca Stringhini
Boston University, US
- Gareth Tyson
Queen Mary University of
London, GB
- Sophie van Der Zee
Erasmus University –
Rotterdam, NL
- Sebastian Wachs
Universität Potsdam, DE
- Victoria Wang
University of Portsmouth, GB
- Jeff Yan
Linköping University, SE
- Savvas Zannettou
Cyprus University of Technology
– Lemesos, CY

