Identifying Unintended Harms of Cybersecurity Countermeasures

Yi Ting Chua
University of Cambridge
yiting.chua@cl.cam.ac.uk

Simon Parkin
University College London
s.parkin@ucl.ac.uk

Matthew Edwards
University of Bristol
matthew.john.edwards@bristol.ac.uk

Daniela Oliveira
University of Florida
daniela@ece.ufl.edu

Stefan Schiffner
University of Luxembourg
stefan.schiffner@uni.lu

Gareth Tyson
Queen Mary University of London
g.tyson@qmul.ac.uk

Alice Hutchings
University of Cambridge
alice.hutchings@cl.cam.ac.uk

***Abstract*—Well-meaning cybersecurity risk owners will deploy countermeasures (technologies or procedures) to manage risks to their services or systems. In some cases, those countermeasures will produce unintended consequences, which must then be addressed. Unintended consequences can potentially induce harm, adversely affecting user behaviour, user inclusion, or the infrastructure itself (including other services or countermeasures). Here we propose a framework for preemptively identifying unintended harms of risk countermeasures in cybersecurity. The framework identifies a series of unintended harms which go beyond technology alone, to consider the cyberphysical and sociotechnical space: displacement, insecure norms, additional costs, misuse, misclassification, amplification, and disruption. We demonstrate our framework through application to the complex, multi-stakeholder challenges associated with the prevention of cyberbullying as an applied example. Our framework aims to generate these consequences, not to paralyze decision-making, so that potential unintended harms can be more thoroughly anticipated and considered in risk management strategies. The framework can support identification and preemptive planning to identify vulnerable populations and preemptively insulate them from harm. There are opportunities to use the framework in coordinating risk management strategy across stakeholders in complex cyberphysical environments.**

***Index Terms*—risk analysis, cybercrime, unintended consequences, unintended harms, countermeasures**

I. Introduction

To manage risks and potential threats to a system of computing devices or an online platform/service, system owners may deploy additional controls — countermeasures — to generally increase security, or to address specific risks. These can range from keeping system software and capabilities up-to-date (e.g., to thwart commodity attacks), to targeted countermeasures to address risks specific to a individual system or situation of concern.

Countermeasures can include technical controls (e.g., advanced verification of user accounts), as well as policies and guidance for users of the system (such as awareness materials, or a declaration of expected conditions of use for a forum or an organisation's IT systems). These countermeasures may be deployed to manage particular risks (e.g., identifying specific language or topics as not being allowed on a social platform), or to raise the minimum level of security within a system to make it safer (e.g., added authentication requirements for accessing a platform).

The deployment of countermeasures is driven by good intentions, to prevent or reduce the harms of particular risks. What is not readily considered is that countermeasures may themselves introduce unintended consequences, be it in crime prevention [1]–[3], physical safety [4], or in IT-security more generally [5]. What is considered even less often is that countermeasures can actually do *harm*, to infrastructure or to some or all of its users. This harm may be as slight as causing disruption and additional security burden for using a system [6], through to negatively impacting whole groups of users such that they are forced away from the system/service or find themselves in a position of increased physical or psychological harm. This paper explores the space of *unintended harms*.

The need to study unintended harms in depth is demonstrated in recent real-world events. One example is the recent deployment of facial recognition in publicly-accessible spaces to augment law enforcement and public control capabilities. This has had the intention of reducing crime and unwanted behaviours, but has (perhaps more so) sparked privacy concerns, particularly in the US [7] and UK [8]. The debate centres on whether such systems are appropriate, given the potential for invasion of privacy and linking of data to other systems. In some cases, the accuracy of the facial recognition systems is insufficient [9], with the consequences of this being potentially harmful to individuals. This is a clear example of where a risk control countermeasure to achieve good can have potentially negative impacts upon people.

Yi Ting Chua and Alice Hutchings are supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M020320/1]. Simon Parkin is supported by grants for the "Gender and IoT" project (UCL Social Science Plus+ scheme, UCL Public Policy, NEXTLEAP Project, and PETRAS IoT Research Hub), and UCL EPSRC Impact Acceleration Account (IAA) 2017-20 project "Investigating Data Security Practices for Small Charities" (EP/R511638/1, KEIF 2017-20 (B.1.3)). Gareth Tyson is supported by the Alan Turing Institute, as part of the "Detecting and understanding harmful content online: a meta tool approach" project (EP/N510129/1)

A. Our contribution

Our contributions are two-fold. We begin by presenting five case studies (Section III) to help highlight commonly observed unintended harms. As our first contribution, we select a range of complex examples, namely: intimate partner abuse, disinformation campaigns, CEO fraud, phishing, and dating fraud, and consider potential interventions that may be applied to each. Within the case studies we also convey how stakeholders acting alone can undo not only their own efforts but also those of others; this points to a need for a shared terminology and strategic thinking between stakeholders. We broadly categorise potential interventions according to whether they are directed towards changing *content* (as outcomes of user behaviour), *users*, or *infrastructure*.

Our second contribution is to provide a framework for conceptualising potential unintended consequences and harms (Section IV). We classify them as imposing additional cost, misuse, insecure norms/complacency, false positives, displacement, amplification, or disruption of other countermeasures. We note that there are often specific populations that are more vulnerable to unintended harms than others (Section V), before applying our framework to the challenge of preventing cyberbullying (Section VI).

The framework has been developed to better understand the potential for unintended consequences. This is important for considering how the harms might be mitigated at the point of designing or deploying countermeasures (Section VII). Another intended purpose is for informing the design of evaluation studies, to ensure that unintended consequences are measured, in addition to the intended outcomes. These all point to potential future applications of the framework, discussed in Section VIII.

II. Background

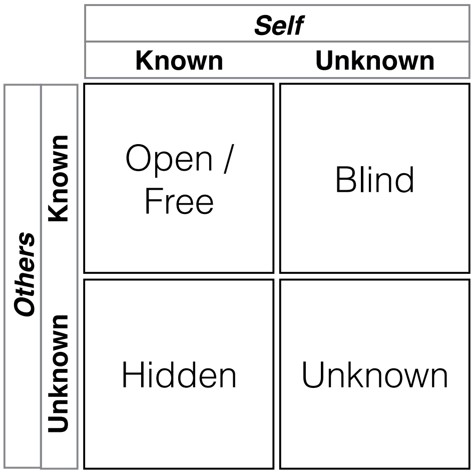
We consider a broad range of risk management situations. We demonstrate co-existing perspectives on risk, where they may be seen as managed in a centralised manner (such as security in an organisation/business, often managed by a security manager or security function), or ultimately must involve a range of stakeholders (including end-users). Both of these perspectives have the potential to influence the larger environment.

A. Countermeasures

We consider that a countermeasure may be deployed with a specific risk in mind. Referring to the ISO 27001 risk treatment framework [10], a risk can be reduced, removed, transferred, or accepted. It should be noted that a risk/system owner may deploy a countermeasure against a *perceived* risk [11]; action may be taken against a risk which a stakeholder believes to be present, or which they anticipate, rather than an existing risk for which there is exhaustive evidence.

A stakeholder may then take action based on limited knowledge about the risk and its impacts upon a system (as in Figure 1). This depiction [12] is adapted from the Johari Window [13]. This is where unintended consequences are critical — if action is taken, it may increase consequences for a particular group, to the point of creating *harms*. A risk/system owner may be best-served by gathering information (or opinions) from other stakeholders in the environment before taking action (increasing the open/free knowledge available to many, as in Figure 1).

Fig. 1. Consideration of limitations to knowledge of risks (and in turn, countermeasures) between one entity and others in the ecosystem - reproduced from [12].



		Self	
		Known	Unknown
Others	Known	Open / Free	Blind
	Unknown	Hidden	Unknown

B. Unintended consequences

Unintended consequences can variously refer to observed phenomena such as ‘knock-on effects’, ‘side-effects’, ‘maladaptive responses’ (e.g., to security awareness campaigns [14]), and so on. Where a countermeasure triggers a string of actions in what appears after the fact to be a sequence, this may be a ‘cascade effect’ [15]; it is then necessary to have a means to map this complexity.

Relevant principles considered in the economics of security help to articulate the characteristics of unintended harms of cybersecurity countermeasures. These include risk dumping, externalities, information asymmetry, and moral hazards.

We regard *risk dumping* [16] as the shifting of risks to entities in the environment who are both unprepared to manage the risk and with whom a negotiation to manage the risk did not happen. We see this commonly in everyday security, in the existence of *workarounds* and *coping strategies* in IT-enabled workplaces [6], as an indirect result of inappropriate risk controls [17].

Externalities refer to the actions of one party creating a positive or negative impact upon another [18]. We consider *harms* as negative side-effects upon another party. Risk management activities are often seen as wholly positive; however, as we demonstrate in our case studies (Section III), some risk management actions can adversely affect other actors in the system and even the risk/system owners themselves. Considering these negative externalities is critical, as they may create additional costs which must be borne by others and not the original risk owner; the original risk owner may in fact be unaware of the burden they have placed on others. This can especially be the case if specific behaviours, users, or technologies are removed as a consequence of a countermeasure, and no longer register within existing risk measurement capabilities (referring to Figure 1, the risk owner becomes 'blind').

Another principle considered in security economics is *information asymmetry* [18], where information or actions are hidden from an entity. This can be critical in supporting those impacted by risk management actions as they may not have the information needed to manage the risk themselves (in the case of risk dumping on end-users, for instance). We consider here that hidden actions can include proactive, well-meaning activities by multiple stakeholders (e.g. [19], [20]). For instance, evidence-gathering by law enforcement may be disrupted if the observed asset is impacted by the risk management activities of others (there are parties unaware of the plans of the risk owner to act upon that asset, as in Figure 1).

This points also to *moral hazard* [21], where a system owner may not take action to, for instance, recover users or user behaviours to their platform which have been forced out, if it does not create any harm *for them*; this can be the case if affected users are disempowered and are unable to register that they have been adversely affected. This is another key aspect of managing unintended harms strategically — a stakeholder may not be incentivised to undo or avoid unintended harms unless they become part of the strategic planning of the platform itself. One example may be if controls prevent some users from using a online social platform, even where the impact on a subset of those users is in effect 'collateral damage'.

These various dynamics characterise *unintended consequences*, pointing to how a risk management action can create *unintended harms*.

C. Unintended harms

We regard *harms* as unintended consequences that have been shifted to another entity in the environment without them being adequately prepared, or able at all to respond to the additional risk they are now burdened with. Similarly, an entity with a stable and safe experience within the managed environment may find themselves moved to another set of circumstances where they no longer enjoy the benefits of the stable environment. Critically, we see two shifts in the management of risks: (1) the current globalised climate of cyber aggression and cyber deception (including potentially commoditised cyber fraud [22]); combining with (2) the increased positioning and proliferation of technologies (and 'cybersecurity' capabilities) in peoples' lives [23]. With these two trends, there is increased possibility of personal harms — physical, cognitive, or psychological — to individuals. Where investigations of risk management have alluded to potential side-effects of countermeasures, as *unintended consequences*, we believe this is the first work to consider development of a dedicated framework for exploring potential *unintended harms*, and prioritising their identification as a first step toward protecting users from being adversely (perhaps irrecoverably) affected.

III. Case Studies

We now describe five *case studies* outlining scenarios involving cyber aggression or cyber deception. For each case study we provide example countermeasures and potential *unintended harms*. The set of countermeasures and associated risks in each scenario is not exhaustive.

A. Intimate partner abuse

Bob and Charlie live together. Charlie is controlling and monitors Bob's behaviour using IoT devices [20]. This includes Bob's smartphone [24]. When suspecting Bob might be visiting friends, Charlie goes on to Twitter and shares aggressive and fabricated posts about Bob [24], [25].

Discard suspect devices. Advice to Bob may be to discard their devices (including smartphones) so that they cannot be tracked [26], [27]. Having *no access* to technology (and potentially with this, online service accounts accessed through those devices) will not help Bob when and if it is necessary to access housing and financial support [28]. In fact, if tech-enabled communication were to continue, it could be a way to monitor and manage Charlie's actions [27], and avoid escalating potential harms. Discarding devices may also *destroy evidence* that could otherwise be used in legal processes/proceedings [27].

Remove harmful content. If Charlie has created offensive online content (or shared intimate content widely online), channels may be created to allow Bob to have the content 'taken down', but this might create an *additional barrier* of Bob needing to find and/or hire a legal professional, working on behalf of Bob [27].

Provide guidance. It may at face value seem useful for experts to *produce advice* for securing personal devices so that people in a similar situation to Bob can control devices in a shared home. This however may require information to be targeted, and available when it is needed [28], as Bob may have limited time alone to act in a climate of abuse. With technology having been used to create harm, Bob may instead fear technology and not wish to use any technology-based solutions [27], so using technology to fix a technology-based problem may simply not be the right approach.

B. Disinformation campaigns

There is a political campaign where Bob and Charlie are both running for governor. A third party, who supports Charlie, conducts a concerted misinformation campaign to spread false information about Bob. This is done predominantly via Facebook and Twitter, and initiated via a network of social media bots which disseminate the material. The overall goal of the campaign is to deceive voters. [29]–[31]

Content removal. This countermeasure generally involves the removal of content, accounts and/or bots [29], [30]. The removal of content may create a '*Streisand effect*', where the request to remove content can draw increased attention to it [29], [32]. In the scenario, removal of content may *backfire* if the third-party presents this as unjust [32], using it as proof of a conspiracy and suppression of 'truth' against them. One potential, topical example of this is in the US far-right movement, which has moved to framing itself at times as a fight against the suppression of white people [33]. Such removal can potentially speed up misinformation diffusion [34], [35].

Account removal. The removal of accounts or content does not address the root cause or motivation for a misinformation campaign; it instead *displaces* a subset of users to other available and more accommodating platforms. For example, in the United States, there is a shift to using platforms such as Gab and Telegram channels for alt-right movement supporters, as a response to bans and removals in mainstream platforms such as Facebook and Twitter [36]. This may in turn facilitate a creation of an 'echo chamber', where individuals surround themselves with information that confirms their own beliefs, opinions, and views and ultimately results in group polarisation [31], [37].

Removal of bots. Although potentially effective [38], this can result in *misclassification*. Misclassification is of increasing concern as social bots' capabilities to generate human-like behaviours are improving [39], [40]. There are two general types of misclassification: false negatives and false positives. *False negatives*, or the misclassifcation of bots as legitimate accounts, can intensify the effects of disinformation since users were found to trust information from bots [41] and bots were found to be more likely to share false information [38]. *False positives*, or misclassification of non-bot accounts as bots, can lead to a perception of censorship among legitimate users [38]. It can also potentially displace users to other platforms (thereby not reducing the risk, but transferring it).

Automated detection algorithms. The development of auto-detection algorithms [30], [38], [39] raises similar potential harms as removal. The goal is to reduce the burden on users in detecting and verifying accuracy and falseness of content and/or accounts [29], [38]. This can, perversely, potentially *reduce users scepticism towards misinformation* [41]. Another unintended consequence of automated detection is *automation-related complacency potential and automation bias*. Complacency refers to poorer detection of malfunctions, while the latter refers to errors made by individuals based on their interactions with imperfect automated decision aids [40].

Fact-checking. Fact-checking [31] may be introduced. This may either incorporate fact-checking as part of content management [30], or encourage users to utilise tools prior to sharing information [41]. With both approaches, an unintended consequence is fostering *a sense of complacency* among users. In the context of Twitter, the effect of fact-checking in changing discourse is mediated by social relationships between users [42], and by content of the fact-check [43]. In addition, the effectiveness of fact-check posts are dependent on the contents level of controversy [43]. Overall, users may potentially utilise services such as Snopes for the purpose of status management, while elites of a community use fact-checking to challenge users of other communities [42]. In this context, fact-checking is used to solidify in-group status and can contribute to group polarisation and fragmentation.

C. CEO fraud

Bob discovers the name and contact details of a major company's CEO. Knowing that the company has a very hierarchical structure, Bob identifies a relevant employee within the finance team: Charlie. Bob sends an email to Charlie, pretending to be the CEO. As emails within the organisation are not cryptographically signed, Bob does an effective job at masquerading as the CEO. The email states that Charlie should immediately pay an invoice, bypassing the usual checks and balances. Due to fear of retribution, Charlie pays the invoice believing the email to be authentic. The money, however, is transferred to Bob's bank account and Charlie is disciplined for his actions.

Behaviour and security culture change. These approaches strive to change working practices, such that employees do not feel compelled to respond to last-minute requests that do not follow correct protocols. This can, however, lead to unintended consequences, most notably a sense of complacency if restrictive technical solutions do not fit with established ways of working (potentially leading to *workarounds* which can themselves be exploited, such as the problem itself of triggering payment transfers from non-corporate email accounts). This is particularly the case as there are a range of challenges in measuring security culture [44], where different groups of employees may be susceptible to attacks (such as CEO Fraud) more than others, without the company becoming aware of it.

Electronic signatures. *Electronic signatures* or alternative forms of blocking spoofed emails (e.g., domain blocking of insecure SMTP servers) may be employed. This involves ensuring that spoofed emails cannot reach employees. This again can result in complacency, as these techniques are rarely 100% effective. In many cases, people will accept unverified emails even in the case of failed signatures. An alternative is to simply *change policies* to prevent email systems from being used to request transfers. Although an effective means, this may negatively *impact productivity* within the company. Furthermore, it is difficult to technically enforce this — consequently, certain employees may breach any such protocol.

Payment authorisation. Another approach is to restructure the organisation, such that employees cannot execute transfers so easily, or transferred be requested with such a lack of checks. For instance, *authentication* could be expected for all transfers. The person performing authorisation could also be *trained* in fraud detection. This may be able to reduce the probability of attack, although it could also create *additional costs* on the employees due to the additional time required for completing every transaction.

Email monitoring. This may be employed to automatically identify cases of fraud. This brings a number of risks, particularly as *false positives* may desensitise users to warnings. This may also trigger *privacy concerns* amongst employees, leading them to disengage from the security mechanisms.

D. Phishing

Bob was recently fired, and subsequently holds bitter resentment towards their former employer. Bob devises a phishing attack against the purchasing department of the company. Bob spoofs the email address of one of the company supplier's contact and sends an email to the department's employees pointing to a web page Bob has set up on a separate website, which prompts visitors for username and password to purportedly get advanced access to new prices for materials and supplies for the next fiscal year. Charlie enters their credentials, which Bob then uses to gain access to the companys materials and supply database. Bob deletes the database causing thousands of pounds in loss to the company.

Automatic detection and filtering. This uses a combination of methods, such as blacklists and machine learning models that use the structural features of email messages (e.g., headers, content, embedded URL) to facilitate detection of phishing messages [45], [46]. Internet Service Providers (ISPs) or email providers block the detected messages from being delivered.

Website takedown. Websites used for phishing may also be taken down [47]. Unintended harms from these approaches include *insecure norms/complacency* due to false negatives; users might acquire a false sense of security, and detection masks the reality that phishing messages and web sites are constantly changing, which makes the problem of detecting all forms of phishing arguably unachievable. Conversely, false positives in filtering messages and taking down of websites can cause users to lose benign, potentially important messages, and for businesses to be wrongly flagged as malicious senders.

User training. Attempts may be made to teach or educate users so they can identify phishing messages [48]–[50]. There are a variety of proposed training approaches, ranging from games [51], [52] to simulated phishing campaigns, especially in corporate environment [49]. Research by Caputo et al. [49] indicates that corporate users can forget training after some time and fall for the same type of attacks after a short period of time. Training can also induce *additional costs* to users, as the nature of end-user training is that they are required to take time to attend training sessions and handle security tasks on top of their primary tasks (for corporate employees) or everyday activities (for home users). This adds to the users 'compliance budget' [53], potentially reducing productivity. For certain users, it might also cause them to become overly sensitive, ignoring legitimate messages they deem suspicious (where this increases false positives).

Internal education or behaviour change activities. Training can potentially *disrupt other countermeasures*, especially when the user receives contradictory advice. For example, training may suggest to be suspicious of emails coming from outside the organisation domain, ignoring that phishing attacks may spoof internal email addresses as well. Another general unintended harm of training or security behaviours is that attackers may learn of it and adapt to work 'outside' of them, such that there is a perpetual 'arms race'.

E. Dating fraud

Charlie is searching for a partner on an online dating site. Charlie encounters Bob, and they hit it off. Unfortunately, Bob lives in Peru and cannot afford to travel to meet Charlie. After a few weeks of intimate conversation, Bob requests $3000 to book a flight and visit Charlie. Once the money has been transferred, Charlie never hears from Bob again.

Verify user identities. As the fraudster has no doubt misrepresented themselves, an obvious countermeasure might involve verifying the identity of dating site users through some technical or administrative means. However, the *cost* of background checks could be prohibitive [54], and crucially, legitimate users may find that the verification process *interferes with their preferred means of self-representation* on a dating site, which could involve some small degree of misrepresentation or selectivity [55]. Where technical implementations rely on connecting social networking accounts, this can expose users to *risks of misuse* by either the dating site (now possessing their public identity, and perhaps an excess of information) or the social networking site (which now possesses potentially sensitive information about their sexuality [56]), and correspondingly increase *safeguarding responsibilities* for these organisations [57].

Close fraudulent accounts. Fraudulent accounts may be closed where these can be identified. Systems for this can be either post-hoc, with the onus on dating site users to report a profile they believe to be fraudulent, or preemptive, with moderators or technical controls screening profiles for markers of suspicious behaviour [54], [58]. Reporting systems can be *abused* by users in redress of personal grievances [59], and so reports must be reviewed by human moderators, a monotonous and thankless job which may create additional risks and harms [60]. Screening mechanisms can also *misfire*, requiring a means of redress, and might be especially *discriminatory* for users from particular backgrounds or locations (e.g., West Africa) [54], unfairly excluding them from an important venue for modern romance.

Press criminal charges. Damaged victims may seek to press criminal charges against a fraudster. The scale of online fraud, and the number of jurisdictional hurdles to clear, make such prosecutions difficult for law enforcement, and can also expose the victim to additional risk of *revictimisation* fraud, in which scammers pretend to be investigators returning the lost money in order to extract further payments from their victims [54].

Provide advice. Quite aside from whether this advice is effective [61], well-intentioned descriptions of 'what to look out for' can provide invaluable advice for crafty fraudsters on how to *disguise themselves*.

F. Summary

Our case studies have illustrated an – albeit limited – range of unintended harms emerging from otherwise well-meaning countermeasures to risks. We discuss the implications of unintended cybersecurity harms upon behaviours, users, and infrastructure in the next section.

IV. Unintended consequences and harms

Based on our analysis of the harms described above (and sundry others), we next strive to create a simple taxonomy, before discussing a simple framework that can be exploited to identify unintended consequences of future countermeasures

A. Taxonomy of Unintended Harms

First, we propose a general taxonomy that captures key types of unintended consequences. We identify seven broad categories:

1) **Displacement**: Crime displacement occurs when crime moves to other locations, times, targets, methods, perpetrators, or offences, as the result of crime prevention initiatives [62]. Examples include alt-right supporters shifting to Gab and Telegram in response to bans and removals on more mainstream social media platforms [36], the surge of new online drug markets following the takedown of Silk Road [63], or phishing sites moving to domains and hosts which are more resistant to takedown efforts [47].
2) **Insecure norms / complacency**: The implementation of countermeasures encourages insecure behaviours, creating the potential for greater harm. Examples include creating a reliance on technical controls [64], and normalising the sharing of personal data for identification purposes.
3) **Additional costs**: Countermeasures can often involve additional costs to particular parties in terms of time or resources. If a cost-benefit analysis has not been performed [65], these costs may even be greater than the original harm. Examples include reporting systems for abuse which pose a burden of manual review for social media companies and their employees [60], and extensive anti-phishing training which places the burden of responsibility on low-level employees.
4) **Misuse**: A countermeasure developed to prevent harm may be intentionally misused by a variety of actors in order to create new harms [66]. Examples include advice for victims being repurposed as training material for perpetrators, reporting systems being used maliciously as a result of personal grievances [59] or competitive business interests [47], and details provided for identity verification purposes being sold to advertisers.
5) **Misclassification**: Technological or administrative systems that create good/bad or allowed/disallowed distinctions will occasionally classify non-malicious content or individuals as malicious. The harm that those affected by misclassification will suffer can be significant if not anticipated. Examples include the "cold start" problem in reputation systems, with legitimate new users being unable to establish credibility to enter a community; users of dating sites being automatically misclassified as scammers and excluded from the dating pool [54]; and stringent identity verification processes preventing individuals without documentation from accessing necessary services.
6) **Amplification**: Interventions can backfire, causing an increase in the behaviour targeted for prevention. Examples include abusers escalating violence when made aware of attempts at disconnecting them from their victims [67], and the well-known 'Streisand effect', where an attempt to take down content causes increased interest in preserving and sharing it [32].
7) **Disrupting other countermeasures**: Countermeasures can interrupt the operation of other, potentially more effective, countermeasures. Examples include devices used in partner abuse being discarded, or abusive online content being taken down, destroying evidence for criminal prosecution [24]. Identity verification schemes can prevent users from protecting themselves from online abuse with anonymity/pseudonymity, and security and safety advice provided for a number of issues can contradict other advice, leading to user confusion [68], [69].

B. A Framework for Unintended Harms

We further developed these unintended harm categories into a framework of questions which may be asked of any (proposed or existing) countermeasure, in order to identify potential negative consequences of deployment upon user behaviours, users, or infrastructure — the framework questions are presented in Table I.

The ordering of the questions in Table I does not imply any ordering of importance, though the final question would ideally be considered after fact-finding efforts to explore questions 1-7. The eighth and final question can be considered as a cross-cutting concern – harm to a particular group might occur through any of the previously-described mechanisms. The explicit consideration of groups allows users of the framework to consciously identify when a countermeasure is shifting risk between stakeholders, something that is especially important when a countermeasure might on net shift harm from less vulnerable groups to more vulnerable groups (as discussed further in Section V).

The questions are deliberately framed to prompt a response, and open enough to prompt consideration of any or all of behaviours, users, and infrastructure, beyond thinking of implications for technological solutions alone.

The framework is intentionally *generative*: it provides prompts for the identification of new possible harms from a countermeasure, without prescribing how the relative likelihood and severity of these harms should be taken into account when developing mitigations, or how they should be weighted against the benefits provided by the countermeasure. A number

TABLE I
FRAMEWORK OF PROBE QUESTIONS FOR EXPLORING CATEGORIES OF UNINTENDED HARMS.

Item	Harm Category	Probe Question
1	Displacement	In what ways might the countermeasure displace harm to others?
2	Insecure norms	In what ways might this countermeasure create insecure norms (especially complacency)?
3	Additional costs	In what ways does the countermeasure burden stakeholders?
4	Misuse	In what ways could the countermeasure be used in attacks?
5	Misclassification	In what ways does incorrect classification cause harm?
6	Amplification	In what ways could the countermeasure amplify harm?
7	Disruption	How might the countermeasure disrupt another countermeasure?
8	ALL	Which groups are more at risk of experiencing harm from the countermeasure?

of existing risk assessment frameworks could potentially be employed for such purposes, such as ISO27001 or the NIST risk management guidelines [70].

Our framework acts as a tool which enables users to consider the following potential outcomes:

(a) Actions introduce whole new classes/types of risk which no existing countermeasures can manage;
(b) Actions exacerbate existing risks/problems which the countermeasures were actually intended to manage;
(c) Actions mask an existing problem.

In this sense, our framework also contributes to the capacity to have a lasting 'memory' of risk management actions - the framework can be applied repeatedly and recursively across harm mitigation proposals.

In such an iterative deployment scenario, the framework can help highlight systemic issues with countermeasure proposals, such as where particular populations would frequently be placed at risk if the proposals were implemented, or certain categories of harm seem to be repeatedly overlooked by those generating countermeasure proposals. These systemic issues can then be addressed by addressing the system producing proposals: do external stakeholders need to be included? Does the scope of proposals need to be extended (e.g., to countermeasures beyond the immediate control of members proposing change, such as legal reform, technical standardisation, or public policy)?

Regarding the involvement of external stakeholders, the deployment of countermeasures related to cybersecurity and cybercrime often involves multiple agencies and stakeholders (e.g. [19], [20]). Stakeholders in mitigating cybercrime can include law enforcement, policymakers, system administrators, and others [71]. A complex approach can result in a failure to assess and manage risks at a higher level, posing challenges in the identification of unintended harms. Our framework then defines terms of reference which can be shared and coordinated across stakeholders.

Unintended harms are not necessarily problematic because they are harms – a risk assessment may rationally conclude that the risk of harm generated by a countermeasure is acceptable given the benefit it will produce. However, because such harms are unintended, and thus unexpected and unknown, they may be excluded from a risk assessment. This can lead to decisions being taken in an under-informed manner (referring to the 'Unknowns' in Fig. 1, pointing to opportunities to become more informed). Our framework aims to generate these consequences, not to paralyze decision-making, so that consequences can be more thoroughly anticipated and considered in risk management strategies.

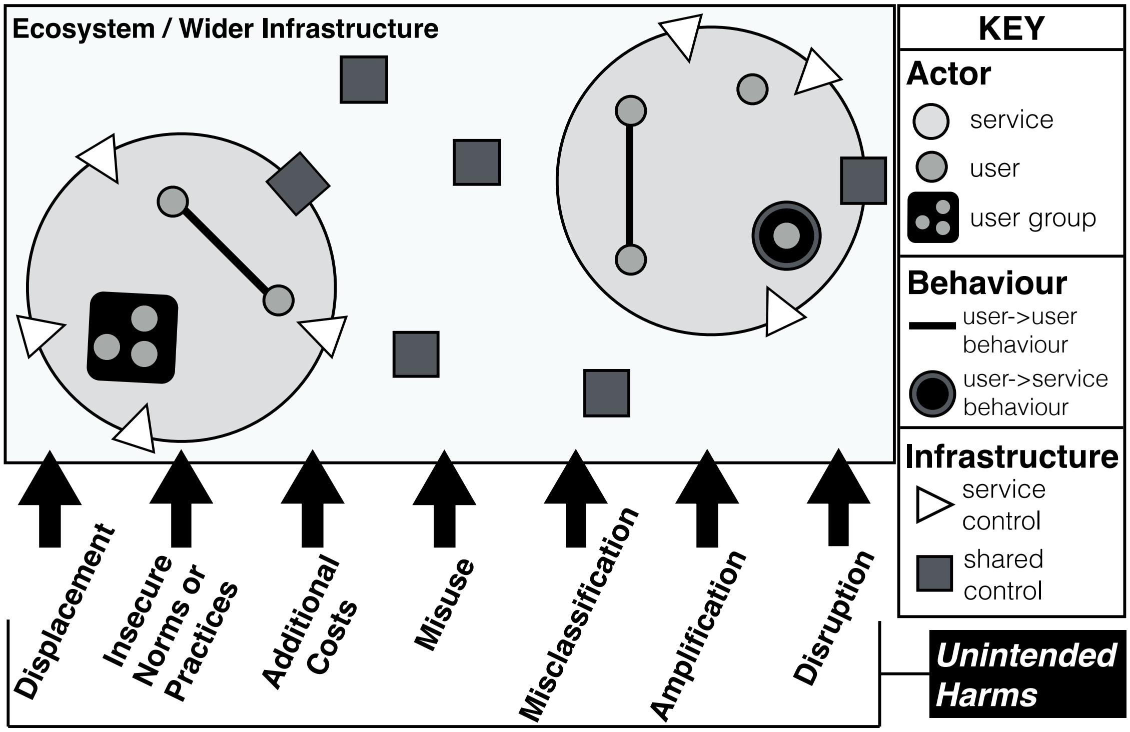
V. VULNERABLE POPULATIONS

When considering the consequences of countermeasures alongside the intended benefits for specific user groups, we may see that some user groups are routinely targeted for protection by countermeasures, while other groups may be left not served and without support. This can include impact upon their use of a service or system (their behaviours), their access (whether and how they are regarded as users of the service/system), and changes to the infrastructure (such as whether particular functions, features, or risk protections are present or remain in place). To be clear, here we consider user groups who are 'collateral damage' of a countermeasure — before deployment of the countermeasure they would have been assumed to be legitimate users who warrant protection and also assumed to not be adversely affected by the countermeasure. This then further motivates the need to preemptively consider how countermeasures may impact distinct users or user groups.

We consider that there can be under-served user groups — vulnerable populations — which:

(i) May be 'forgotten', especially if they are (inadvertently) removed from the service or their behaviours are restricted. Here, design accommodations could be made in advance to avoid this. This also relates to the masking of problems, as in potential outcome (c) in Section IV.
(ii) May not have the access or capabilities needed to make use of provisioned controls. Hence, they may need access to more appropriate, *alternative controls*.
(iii) May be affected by a combination of harms. That is, it is not necessarily always the case that only one harm affects a user or user group at any time. The more distinct harms which affect a group, the more the more they should be considered as 'vulnerable' or in need of concerted assistance. This relates to potential outcome (b) (c) in Section IV.

Fig. 2. The potential unintended harms of a countermeasure and elements of a system which may be affected by those harms.



(iv) May be affected by one or more harms continually/regularly over time (requiring sustainable solutions, representation, and monitoring/engagement to match sustained challenges).

Referring to Figure 2, for (i) above, a user or user group may be taken out of a service environment and is then reliant on existing (shared) controls in the wider ecosystem (such as advocacy groups or basic protections provided in technology, *unless* specific support is provided.

For (ii) above, a user or user group may remain within a service environment but not be able to use service-specific or shared controls — an example would be a presumption that users (and especially, older adults) are comfortable and experienced enough with technology to conduct online banking on a secured website (if the capacity to conduct banking in-person at a branch is disrupted). An appropriate control for impacted users may be unavailable, unreachable, or may not yet exist.

For (iii) above, any combination of harms (as in Figure 2) may impact a user or user group, for instance a lack of technical skills combined with technologies which may be misused by a skilled attacker/perpetrator.

For (iv) above, this would require a capacity to measure when users, behaviours, groups or controls cross boundaries between individual services and the wider ecosystem; an example would be when employees in companies are provided with IT support as part of their work, but then are not supported to the same level to be secure online when they retire [72].

A. Vulnerable user groups

Prior work has examined whether security behaviour interventions have a non-uniform effect to reduce harmful intentions across different groups of (potentially malicious) users [73], referring to this as the ‘differential effects’ of information security countermeasures. We explore instead the differential effects upon users who *the infrastructure owners would not want to be in a state which is any more malicious or un-secure than it currently is.*

Here we consider a range of vulnerable user groups, where this set is far from intended to be exhaustive, but instead to provide a sense of how some groups in the risk-managed environment may be disadvantaged by countermeasures more than others. We also demonstrate here how *these groups are disadvantaged while others may be unaffected or continue to prosper* (this being our definition of a ‘vulnerable population’). We then also highlight how a ‘vulnerable population’ is ‘hidden’ from view compared to other groups, and hence more adversely affected (and potentially failing to register in the awareness of the risk/system owner). For these reasons, harms for distinct populations can include *risk dumping*; *unpredictable/destabilised cyber-physical environment*, or; *masking of the risks* from the view of the risk owner.

Note also that a population or user group considered ‘vulnerable’ in one service domain may not be in another. For example, a user group inadvertently prohibited from accessing a specific service may not lack the technical skills to find another comparable service, but nonetheless could have been spared the additional cost if the events leading to their exclusion were preempted. Referring to concepts represented across Figures 1 and 2, this can be a combination of users and their activities being inadvertently affected, or the burden on them no longer being known to the (now former) risk owner.

- **Older adults.** May have had fewer opportunities to habituate use of technologies. However, older adults may also have a heightened sense of risks. These users may be 'hidden' from the view of risk owners, as they potentially interact with technology support rarely – for instance, going to a store to buy a new device relatively rarely [72], delegating the security of their device to a paid 'IT person', or otherwise never or rarely interacting with technology [74].
- **Small businesses.** Smaller businesses may have less resources available to invest in automated security solutions, and be less likely to have a dedicated security function to manage threats [75]. However, smaller businesses may also have individual staff members with an increased knowledge of how various parts of the business operate and use IT, compared to employees in larger organisations where IT — and cybersecurity — may be directly managed by a designated security function. They may also delegate security to an expert IT-security provider company [76].
- **Survivors/victims of tech-abuse.** May be controlled or monitored both physically and through devices and online services. Opportunities to configure or modify devices may be limited, and there are potential implications if a perpetrator discovers interactions with a device. However, there are frontline services skilled in addressing domestic and intimate partner abuse. This group may be 'hidden' from view as the configuration of consumer computing environments is often assumed to be agreed between all members of a shared living space.

VI. Application to New Scenario: Cyberbullying

Having outlined our framework, we now apply it to a new case study of cyberbullying:

Jill is a seventh grader. For the past year, Joey and other classmates have been leaving aggressive comments such as "You're so stupid" and "You smell" on Jill's Facebook profile. Joey and other classmates also found out about Jill's Snapchat account and have been sending disturbing and threatening images to Jill every day. [77]

Various stakeholders have developed and implemented a range of cyberbullying countermeasures (e.g., [19], [78], [79]). Countermeasures in this space are not necessarily risk-free, and merit careful assessment given that one of the target audience groups is young users [80], [81].

In some instances, the unintended harms of the countermeasures outweigh the benefits. For example, encouraging children to include false information with accurate information can interfere with automated detection and filtering systems [82]. In the long run, these countermeasures have the power to change the targeted behaviours and how the next generation interacts with technology. Therefore, it is necessary to assess all possibilities prior to implementing any countermeasure.

As it becomes easier to connect with others via the Internet and social media, there is a rise in prevalence of cyberbullying and online harassment among teenagers and young adults [83]–[85]. Cyberbullying victimisation is shown to correlate with an array of negative consequences. For example, Hinduja and Patchin [86] found that individuals who were cyberbullied were more likely to report engagement in offline problem behaviours such as running away from home or carrying a weapon. Females were also more likely to be victims of cyberbullying [84], [87]. Psychologically, victims of cyberbullying and school-based bullying were more likely to report suicidal ideation compared to those who did not experience any type of bullying [88].

These negative outcomes lead to the introduction of a multitude of countermeasures. To illustrate the applicability of the proposed framework, we will focus on two common countermeasures for cyberbullying – educational and training, and privacy control and management – and identify potential unintended consequences associated with each countermeasure.

A. Education and training — unintended harms

Education and training is frequently recommended to various stakeholders, such as teenagers, parents, and educators [19], [78], [79], [81]. The purpose is to establish basic knowledge on cyberbullying and appropriate online behaviours, and to communicate the consequences of cyberbullying [81], [89], [90].

For parents, teachers and school administrators, the goals of education and training differ. Rather than establishing basic knowledge, these programs focus on proper responses to and prevention of cyberbullying [89], [90]. There are also programs that place an emphasis on building protective factors, such as positive school climate [91] and resilience [92], to minimise the negative impacts of cyberbullying.

Using our framework (Section IV), we outline the unintended harms of *education and training*:

- **Displacement.** There are two possible types of displacement. First, cyberbullies may adapt their behaviours to circumvent detection. For example, teenagers are advised to disregard minor teasing and not engage with aggressors [78], [79]. Such advice can potentially result in cyberbullies switching to this strategy compared to more well-known and problematic cyberbullying behaviours (e.g. sending threatening text and messages). Second, there is the possibility of migration to social media platforms that are more lax and provide more freedom to users. For example, in 2013, teenagers started migrating away from Facebook to other social media platforms such as

Instagram where cyberbullying is more prevalent [85], [93].

- **Insecure norms/complacency.** This countermeasure might create a false sense of security among stakeholders. Despite a large number of available resources and educational programs, there is very little empirical evidence on their effectiveness [81], [90], [94]. For instance, most victims of cyberbullying do not disclose to adults [84] or utilise the block function of online communication tools [95]. Although these findings are dated at this point, they highlight the need to assess if and which education and training programs are effective.
- **Additional costs.** This countermeasure places extra burden for stakeholders in terms of the effort, resources and time needed to develop and implement these programs. Teachers and educators need to allocate time to attend training sessions and/or become trainers for other staff in schools [19], [96]. For school administrators, the burden lies in coordinating and incorporating these programs into existing curriculum and community involvement [19], [89]. In fact, recommended practices often emphasise the role of schools in initiating education and training programs [19], [79].
- **Misuse.** The knowledge and information made available through education and training, especially school-wide programs, may potentially be used by perpetrators. Studies have shown that being victims of cyberbullying correlate with future engagement in cyberbullying behaviours [77], [84]. Individuals who attend these programs would now have knowledge on techniques for cyberbullying.
- **Misclassification.** With education and training, incorrect classification arises when definitions of cyberbullying (which lack consensus [81]) become broad enough or so easily misinterpreted that ordinary childhood interactions become labelled – both mislabelled ‘bullies’ and their ‘victims’ might suffer as a result of education programs that identify them as part of a group that needs either censuring or safeguarding. Misidentified bullies can become scapegoats for the misbehaviour of peers, and victims can suffer additional (or actual) bullying as a result of being labelled [97].
- **Amplification.** Education and training may increase the occurrence of victim blaming. Currently, victim blaming is present in pre-teens’ and teenagers’ discourse on cyberbullying where responsibility is placed on victims because of their actions, or lack thereof [98]. To illustrate, consider a form of direct bullying where the cyberbully sends an email with a malicious attachment [77]. The recipient who opens the email may be blamed, as education programs specifically warn individuals not to open suspicious emails [80]. In this sense, the implementation of education and training programs can amplify victim blaming by placing even more responsibility on victims in recognising cyberbullying behaviours and/or following proper use of technology.
- **Disrupting other countermeasures.** With a multi-stakeholder approach on education and training [19], [79], [81], the likelihood of confusion and contradictory information is high. The current lack of consensus on the definition of cyberbullying [81] means that students may potentially be receiving different lists of cyberbullying behaviours from their parents, teachers, and social media platforms.
- **Vulnerable population.** With education and training, there are two potential groups that are at higher risks of experiencing unintended consequences and harm. The first group is the victims of cyberbullying. The implementation of educational program and training may worsen victim blaming among pre-teens and teenagers [98]. The second group includes pre-teens and adolescents who experience physical isolation and rely on online communities for social support. For example, adolescents diagnosed with cancer rely on online forums to exchange experiences and cope with emotions [99]. The frequent use of online communities meant that these individuals are more likely to be impacted by unintended harms of this countermeasure.

B. *Privacy control and management — unintended harms*

This category of countermeasure focuses on the availability and accessibility of personal sensitive information of teenagers in the cyberspace. This countermeasure tends to target pre-teens and teenagers where they are advised to reflect before sharing any information and learn about privacy setting controls for devices, applications, and social media platforms [78], [79].

Beyond pre-teens and teenagers, this countermeasure applies to parents as well. Parents are advised to be directly involved in privacy control and management by searching for their child’s name and making an information removal request for unwanted materials [90]. Parents have also suggested to their children to blend false information when sharing personal information online as a privacy management technique [82], or rely on applications that promote online safety via monitoring [100].

Using the proposed framework, the following section discusses the unintended consequences and harms with *privacy control and management*:

- **Displacement.** There are two potential types of displacement. First, it may encourage migration to other types of platforms with easy-to-use privacy control setting. This may become a pull factor since some pre-teens reported unawareness or lack of understanding on privacy settings on sites such as Facebook [82]. An example is the migration to Snapchat [101] that advertise straightforward privacy features such as deletion of content [102]. Second, this countermeasure may displace harms to individuals without proper privacy settings and controls. This group of individuals may become easy targets.
- **Insecure norms/complacency.** Privacy control and management may foster a sense of complacency among

stakeholders. With applications such as Snapchat, users may be more comfortable with sending less safe materials as the content is deleted once opened [102]. For other applications such as Facebook, users may rely on the default privacy settings [82].

- **Additional costs.** Privacy control and management brings additional costs and efforts to pre-teens and teenagers. They need to dedicate time for learning about privacy setting controls across devices, applications and platforms [78], [79]. The total costs are greater if an individual has multiple accounts and devices, which is quite common among pre-teens and teenagers. In the United States, a large proportion of teenagers have more than one account on social networking sites [103]. This means that they may either use the default setting or do not take the time to learn how to change them to more private settings.
- **Misuse.** Privacy control and management can potentially be used for cyberbullying. First, cyberbullies can create multiple fake accounts with a mixture of accurate and false information [77]. This can overwhelm individuals as they monitor their online presence. Second, cyberbullies can isolate targeted individuals by requesting the removal of these individuals' legitimate accounts and accurate information available via search engines.
- **Misclassification.** With privacy control and management, the likelihood of incorrect classification is low because the main purpose of this countermeasure is to minimise and limit personal and sensitive information that is available and accessible online. [78], [80]. One scenario where there may be unintended consequence is parental privacy control and management [90]. There may be discrepancies between what parents and teenagers deem as unacceptable and/or inappropriate information. Such discrepancies may result in teenagers experiencing breaches of their online privacy.
- **Amplification.** Privacy control and management may result in the Streisand effect [32]. When teenagers manage their online presence by requesting information to be taken down, it may potentially draw more attention to it among their peers.
- **Disrupting other countermeasures.** This countermeasure, especially with the practice of mixing accurate and false information [82], may interfere with countermeasures that rely on automated detection and filtering [94], [104]–[107]. Purposeful inclusion of false information may result in misclassification and/or biases in these algorithms and programs.
- **Vulnerable population.** A vulnerable population that is more at risk for experiencing unintended consequences and harms is young users. There is some evidence showing they engage in safe practices, such as adjusting privacy settings on social mediate sites, but at the same time, there are individuals who seem to be unaware of such settings [82]. The small sub-set of young users is of concern and may be more vulnerable as their peers adopt safer practices.

VII. Opportunities — building on outcomes

The outcomes of applying the framework can be used to identify interventions when signs of risks emerge in a socio-technical system, such as when online relationship activity begins to include unhealthy behaviours [108]. In a similar vein, the framework could be used as a structure for *pre-mortems* [109], where this would be an exercise to identify actions which might be taken in the current moment which may contribute to a process failing in the future. This prompts consideration of what action could also be taken ahead of that time to insulate the process (for instance, a service or technology) and against that failure, in this case weaving unintended harms into the discussion.

Looking specifically at service and technology design, it can be possible to feed the outcomes of applying the framework in efforts to, for instance, Design Against Crime [110]. The analysis in the cyberbullying scenario identified many signals and events to either look out for or avoid on social media or online communication platforms, where crucially this may be happening against a backdrop of *providers wanting to encourage active and positive use* of those services. That is, the unintended harms can be considered against positive service attributes which can also be engineered to minimise potential harms (such as providing advice and support to young people on social media platforms, which they can take with them should they deliberately or inadvertently find themselves online but outside of that platform at any point, e.g., on an unfamiliar chat-room or forum).

The literature on situational crime prevention framework [111] and problem-oriented policing [112] emphasises the need to consider displacement in the selection and assessment of crime prevention strategies, discussing how to account and measure for it. This goes to demonstrate the feasibility of a risk management and assessment framework in designing and deploying countermeasures — here we approach cyberphysical crime challenges, but also aim to develop the makings of a tool which can be used by a range of stakeholders (not just law enforcement) toward a coordinated approach in a complex, multi-party service/technology ecosystem. To that end, the framework may also complement existing multi-stakeholder capabilities, such as Multi-Agency Risk Assessment Conferences (MARAC) [113] which are arranged to manage cases of domestic abuse.

Future work will then investigate the compatibility of our proposed framework with existing risk assessment and management literature. For instance, the proposed framework is compatible with the notion of residual risk, or risks that remain after appropriate security controls for risk reduction are in place [114]. With the proposed framework, an assessment of residual risks will also include examining risks posed by nominated security controls.

Alongside these efforts, a broadening of the security economics principles visited in Section II can be developed to support a structured cost-benefit analysis of risk management strategies. To consider generally that all countermeasures carry some kind of adverse side effects, the implementation of countermeasures should be approached in a manner which is conscious of these effects and facilitates a trade-off of side effects with benefits. Within this is a need to consider decision-maker *preferences*, where harms should not be induced upon user groups who are known to be unprepared or unsupported (Section V).

VIII. Conclusions

The paper has studied the process of deploying cyber-countermeasures, as well potential unintended consequences. Unintended consequences, and unintended harms, are of increasing importance for two reasons. Firstly, unintended harms can potentially have far-reaching impacts in current society where technology and the Internet is highly incorporated into our daily lifestyles. The countermeasure for one aspect of socio-technical interaction may alter the norms with other aspects of the interaction. The integration of technology into our daily lives masks the possibility that the impacts of a countermeasure can be hidden from view unless they are deliberately and proactively explored.

Second, as illustrated in our case studies, the deployment of countermeasures related to cybersecurity and cybercrime often involved multiple agencies and stakeholders. The need for efforts from stakeholders adds another layer of complexity when assessing unintended harms and consequences, especially in circumstances if there is a lack of coordination and communication between agencies and stakeholders managing the same issue.

Both reasons highlight the need for a strategic approach to uncover cyberphysical and socio-technical implications of any one intervention. Our framework illustrates the capacity to consider unintended harms when assessing existing countermeasures, as well as its potential application to real-world scenarios. All in all, our framework provides guidance and starting points for stakeholders to incorporate the discussion of unintended harms as part of broader risk management strategies, with the greater aim of supporting cybersecurity practices which act to limit unintended harms to society and its constituent user groups.

Acknowledgment

The authors would like to thank Schloss Dagstuhl and the organisers of Dagstuhl Seminar 19302 ("Cybersafety Threats - from Deception to Aggression") for supporting this work, and attendees of the seminar for their feedback.

References

[1] Peter N Grabosky. Unintended consequences of crime prevention. In *In Crime Prevention Studies*. Citeseer, 1996.

[2] Joan McCord. Cures that harm: Unanticipated outcomes of crime prevention programs. *The Annals of the American Academy of Political and Social Science*, 587(1):16–30, 2003.

[3] Brandon C Welsh and David P Farrington. Toward an evidence-based approach to preventing crime. *The ANNALS of the American Academy of Political and Social Science*, 578(1):158–173, 2001.

[4] Sidney Dekker. *The field guide to understanding'human error'*. CRC press, 2017.

[5] Shari Pfleeger and Robert Cunningham. Why measuring security is hard. *IEEE Security & Privacy*, 8(4):46–54, 2010.

[6] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):41–46, 1999.

[7] Dave Lee. San francisco is first us city to ban facial recognition, 2019. Accessed: 08.15.2019.

[8] Zoe Kleinman. Facial recognition in king's cross prompts call for new laws, 2019. Accessed: 08.15.2019.

[9] BBC News. 2,000 wrongly matched with possible criminals at champions league, 2018. Accessed: 08.16.2019.

[10] International Organization for Standardization. *ISO/IEC 27001: 2013: Information Technology–Security Techniques–Information Security Management Systems–Requirements*. International Organization for Standardization, 2013.

[11] John Adams. Risk, university college london press. *London, UK*, 1995.

[12] Michael J Massie and A Terry Morris. Risk acceptance personality paradigm: How we view what we don't know we don't know. *American Institute of Aeronautics and Astronautics*, 1-18, 2011.

[13] Joseph Luft and Harry Ingham. The johari window. *Human relations training news*, 5(1):6–7, 1961.

[14] Geordie Stewart and David Lacey. Death by a thousand facts: Criticising the technocratic approach to information security awareness. *Information Management & Computer Security*, 20(1):29–38, 2012.

[15] Maria Grazia Porcedda and David S Wall. Cascade and chain effects in big data cybercrime: Lessons from the talktalk hack. In *Please cite as: Porcedda, MG and Wall, DS (2019) Cascade and Chain Effects in Big Data Cybercrime: Lessons from the TalkTalk hack, proceedings of WACCO 2019: 1st Workshop on Attackers and Cyber-Crime Operations, Held Jointly with IEEE EuroS&P*, 2019.

[16] Ross Anderson. *Security engineering*. John Wiley & Sons, 2008.

[17] I Kirlappos, S Parkin, and MA Sasse. Learning from shadow security: Why understanding non-compliant behaviors provides the basis for effective security. In *USEC14 Workshop on Usable Security*, pages 1–10, 2014.

[18] Ross Anderson and Tyler Moore. Information security economics–and beyond. In *Annual International Cryptology Conference*, pages 68–91. Springer, 2007.

[19] Michael A Couvillon and Vessela Ilieva. Recommended practices: A review of schoolwide preventative programs and strategies on cyberbullying. *Preventing School Failure: Alternative Education for Children and Youth*, 55(2):96–101, 2011.

[20] Isabel Lopez-Neira, Trupti Patel, Simon Parkin, George Danezis, and Leonie Tanczer. 'Internet of Things': How abuse is getting smarter. *Safe–The Domestic Abuse Quarterly*, 63:22–26, 2019.

[21] Lawrence A Gordon, Martin P Loeb, and Tashfeen Sohail. A framework for using insurance for cyber-risk management. *Communications of the ACM*, 46(3):81–85, 2003.

[22] Michael Levi, Alan Doig, Rajeev Gundur, David Wall, and Matthew Williams. Cyberfraud and the implications for effective risk-based responses: themes from uk research. *Crime, Law and Social Change*, 67(1):77–96, 2017.

[23] Camille L Ryan and Jamie M Lewis. *Computer and Internet use in the United States: 2015*. US Department of Commerce, Economics and Statistics Administration, US , 2017.

[24] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 667. ACM, 2018.

[25] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in south asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 2. ACM, 2019.

[26] Martin Emms, Budi Arief, and Aad van Moorsel. Electronic footprints in the sand: Technologies for assisting domestic violence survivors. In *Annual Privacy Forum*, pages 203–214. Springer, 2012.

[27] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: a qualitative analysis with multiple stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW) Vol*, 1, 2017.

[28] Tech vs Abuse (Comic Relief). Tech vs abuse: Design principles, 2019.

[29] Twitter Inc. Retrospective Review: Twitter, Inc. and the 2018 Midterm Elections in the United States. Technical report, Twitter,Inc., February 2019. https://blog.twitter.com/content/dam/blog-twitter/official/en_us/company/2019/2018-retrospective-review.pdf.

[30] Facebook Newsroom. What is facebook doing to address the challenges it faces? https://newsroom.fb.com/news/2019/02/addressing-challenges/, February 2019. Accessed: 07.28.2019.

[31] Gulizar Haciyakupoglu, Jennifer Hui, V. S. Suguna, Dymples Leong, and Muhammad Faizal Bin Abdul Rahman. Countering fake news: A survey of recent global initiatives. Technical report, Nanyang Technological University, 3 2018.

[32] Sue Curry Jansen and Brian Martin. The streisand effect and censorship backfire. *International Journal of Communication*, 9:656–671, 2018.

[33] Tammy Castle, Lars Kristiansen, and Lantz Shifflett. White racial activism and paper terrorism: A case study in far-right propaganda. *Deviant Behavior*, pages 1–16, 2018.

[34] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[35] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[36] Ephrat Livni. Twitter, Facebook, and Insta bans send the alt-right to Gab and Telegram. https://qz.com/1617824/twitter-facebook-bans-send-alt-right-to-gab-and-telegram/, May 2019. Accessed: 05.12.2019.

[37] Cass R Sunstein. *Republic.com 2.0*. Princeton University Press, Princeton, NJ, 2007.

[38] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(4787):1–9, 07 2018.

[39] Emilio Ferrara, Onur Varol, Clayton David, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 07 2016.

[40] Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 06 2010.

[41] Joanna M. Burkhardt. *Combating Fake News in the Digital Age*, volume 53 of *8*. American Library Association, 2017.

[42] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 187–196, 01 2014.

[43] R. Kelly Garrett, Erik C. Nisbet, and Emily K. Lynch. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naive theory. *Journal of Communication*, 63(4):617–637, 2013.

[44] ENISA. Cybersecurity culture guidelines: Behavioural aspects of cybersecurity, 2019.

[45] A Almomani, B B Gupta, S Atawneh, A Meulenberg, and E Almomani. A Survey of Phishing Email Filtering Techniques. *IEEE Communications Surveys Tutorials*, 15(4):2070–2090, 2013.

[46] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Identifying suspicious urls: An application of large-scale online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 681–688, New York, NY, USA, 2009. ACM.

[47] Alice Hutchings, Richard Clayton, and Ross Anderson. Taking down websites to prevent crime. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10. IEEE, 2016.

[48] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit*, eCrime '07, pages 70–81, New York, NY, USA, 2007. ACM.

[49] Deanna Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M.Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *Security & Privacy, IEEE*, 12:28–38, 01 2014.

[50] Olga A. Zielinska, Rucha Tembe, Kyung Wha Hong, Xi Ge, Emerson Murphy-Hill, and Christopher B. Mayhorn. One phish, two phish, how to avoid the internet phish: Analysis of training strategies to detect phishing emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1):1466–1470, 2014.

[51] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS)*, volume 229, pages 88–99, 01 2007.

[52] Nalin Arachchilage and Steve Love. A game design framework for avoiding phishing attacks. *Computers in Human Behavior*, 29:706714, 05 2013.

[53] Adam Beautement, M Angela Sasse, and Mike Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 47–58. ACM, 2009.

[54] Aunshul Rege. What's love got to do with it? exploring online dating scams and identity fraud. *International Journal of Cyber Criminology*, 3(2), 2009.

[55] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication*, 11(2):415–441, 2006.

[56] Jeremy Birnholtz, Colin Fitzpatrick, Mark Handel, and Jed R Brubaker. Identity, identification and identifiability: The language of self-presentation on a location-based mobile dating app. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 3–12. ACM, 2014.

[57] Kath Albury, Jean Burgess, Ben Light, Kane Race, and Rowan Wilken. Data cultures of mobile dating and hook-up apps: Emerging issues for critical social science research. *Big Data & Society*, 4(2):1–11, 2017.

[58] Guillermo Suarez-Tangil, Matthew Edwards, Claudia Peersman, Gianluca Stringhini, Awais Rashid, and Monica Whitty. Automatically dismantling online dating fraud. *IEEE Transactions on Information Forensics and Security*, 2019.

[59] Jessica Anderson, Matthew Stender, Sarah Myers West, and Jillian C York. Unfriending censorship. Technical report, Onlinecensorship.org, 2016.

[60] Sarah T Roberts. *Behind the screen: The hidden digital labor of commercial content moderation*. PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[61] Monica T Whitty. Who can spot an online romance scam? *Journal of Financial Crime*, 26(2):623–633, 2019.

[62] R G Smith, N Wolanin, and G Worthington. *Trends & Issues in Crime and Criminal Justice No. 243: e-Crime solutions and crime displacement*. Canberra: Australian Institute of Criminology, 2003.

[63] Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 33–48, 2015.

[64] Alexios Mylonas, Anastasia Kastania, and Dimitris Gritzalis. Delegate the smartphone user? security awareness in smartphone platforms. *Computers & Security*, 34:47–66, 2013.

[65] Borka Jerman-Blažič et al. An economic modelling approach to information security risk management. *International Journal of Information Management*, 28(5):413–422, 2008.

[66] Mario Silic. Dual-use open source security software in organizations–dilemma: help or hinder? *Computers & Security*, 39:386–395, 2013.

[67] Cindy Southworth, Shawndell Dawson, Cynthia Fraser, and Sarah Tucker. A high-tech twist on abuse: Technology, intimate partner stalking, and advocacy. *Violence Against Women Online Resources*, 2005.

[68] Hazel Murray and David Malone. Evaluating password advice. In *2017 28th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE, 2017.

[69] Stacy Tye-Williams and Kathleen J Krone. Identifying and re-imagining the paradox of workplace bullying advice. *Journal of Applied Communication Research*, 45(2):218–235, 2017.

[70] Gary Stoneburner, Alice Y Goguen, and Alexis Feringa. NIST SP 800-30 risk management guide for information technology systems, 2002.

[71] David Maimon and Eric R Louderback. Cyber-dependent crimes: an interdisciplinary review. *Annual Review of Criminology*, 2:191–216, 2019.

[72] Simon Parkin, Elissa M Redmiles, Lynne Coventry, and M Angela Sasse. Security when it is welcome: Exploring device purchase as an opportune moment for security behavior change. In *Proceedings*

of the Workshop on Usable Security and Privacy (USEC'19). Internet Society, 2019.

[73] John D'Arcy and Anat Hovav. Does one size fit all? examining the differential effects of is security countermeasures. *Journal of Business Ethics*, 89(1):59, Aug 2008.

[74] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. Privacy and security threat models and mitigation strategies of older adults. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, Santa Clara, CA, August 2019. USENIX Association.

[75] Simon Parkin, Andrew Fielder, and Alex Ashby. Pragmatic security: modelling it security management responsibilities for sme archetypes. In *Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats*, pages 69–80. ACM, 2016.

[76] E. Osborn and A. Simpson. Risk and the small-scale cyber security decision making dialoguea uk case study. *The Computer Journal*, 61(4):472–495, 2018.

[77] Heidi Vandebosch and Katrien Van Cleemput. Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society*, 11(8):1349–1371, 2009.

[78] Facebook. Empower teens, 2016. Accessed: 08.12.2019.

[79] Sameer Hinduja and Justin Patchin. *Cyberbullying: Identification, Prevention, & Response*. Cyberbullying Research Center, 2018.

[80] Sameer Hinduja and Justin W. Patchin. Preventing cyberbullying: Top ten tips for teens. https://cyberbullying.org/Top-Ten-Tips-Teens-Prevention.pdf, June 2018. Accessed: 08.12.2019.

[81] Russell A Sabella, Justin W Patchin, and Sameer Hinduja. Cyberbullying myths and realities. *Computers in Human Behavior*, 29(6):2703–2711, 2013.

[82] Katie Davis and Carrie James. Tweens' conceptions of privacy online: implications for educators. *Learning, Media and Technology*, 38(1):4–25, 2013.

[83] Justin W Patchin and Sameer Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2):148–169, 2006.

[84] Qing Li. New bottle but old wine: A research of cyberbullying in schools. *Computers in Human Behavior*, 23(4):1777–1791, 2007.

[85] Taylor Lorenz. Teens are being bullied 'Constantly' on instagram. https://www.theatlantic.com/technology/archive/2018/10/teens-face-relentless-bullying-instagram/572164/, October 2018. Accessed: 08.12.2019.

[86] Sameer Hinduja and Justin W Patchin. Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, 6(3):89–112, 2007.

[87] Sally Ho. Girls report three times more online harassment than boys amid rise in cyberbullying. https://time.com/5636059/girls-online-harassment-cyberbullying-report/, July 2019. Accessed: 08.12.2019.

[88] Sameer Hinduja and Justin W Patchin. Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, 18(3):333–346, 2019.

[89] Catherine D Marcum and George E Higgins. Examining the effectiveness of academic scholarship on the fight against cyberbullying and cyberstalking. *American Journal of Criminal Justice*, pages 1–11, 2019.

[90] John Snakenborg, Richard Van Acker, and Robert A Gable. Cyberbullying: Prevention and intervention to protect our children and youth. *Preventing School Failure: Alternative Education for Children and Youth*, 55(2):88–95, 2011.

[91] Sameer Hinduja and Justin W. Patchin. Preventing cyberbullying: Top ten tips for educators. https://cyberbullying.org/Top-Ten-Tips-Educators-Cyberbullying-Prevention.pdf, June 2018. Accessed: 08.12.2019.

[92] Sameer Hinduja and Justin W Patchin. Cultivating youth resilience to prevent bullying and cyberbullying victimization. *Child Abuse & Neglect*, 73:51–62, 2017.

[93] Emily Siner. Facebook takes on cyberbullies as more teens leave site. https://www.npr.org/sections/alltechconsidered/2013/11/07/243710885/facebook-takes-on-cyberbullies-as-more-teens-leave-facebook, November 2013. Accessed: 08.12.2019.

[94] Janneke M van der Zwaan, Virginia Dignum, Catholijn M Jonker, and Simone van der Hof. On technology against cyberbullying. In *Responsible Innovation 1*, pages 369–392. Springer, 2014.

[95] Jaana Juvonen and Elisheva F Gross. Extending the school grounds?bullying experiences in cyberspace. *Journal of School Health*, 78(9):496–505, 2008.

[96] Massachusetts Aggression Reduction Center. Preventing cyberbullying: Top ten tips for educators. https://www.marccenter.org/educator, 2018. Accessed: 08.14.2019.

[97] Dieter Zapf and Stale Einarsen. Individual antecedents of bullying: Victims and perpetrators. *Bullying and emotional abuse in the workplace. International perspectives in research and practice*, 165:183, 2003.

[98] Online Media Law. Youth perceptions of risk, law and criminality on social media (press briefing). https://www.onlinemedialawuk.com/blog/briefing, May 2019. Accessed: 08.13.2019.

[99] Brad Love, Brittani Crook, Charee M Thompson, Sarah Zaitchik, Jessica Knapp, Leah LeFebvre, Barbara Jones, Erin Donovan-Kicken, Emily Eargle, and Ruth Rechis. Exploring psychosocial support online: a content analysis of messages in an adolescent and young adult cancer community. *Cyberpsychology, Behavior, and Social Networking*, 15(10):555–559, 2012.

[100] Pamela Wisniewski. The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience? *IEEE Security & Privacy*, 16(2):86–90, 2018.

[101] Olivia Solon. Teens are abandoning facebook in dramatic numbers, study finds, June 2018. Accessed: 08.15.2019.

[102] Privacy centre - our privacy principles, 2019. Accessed: 08.15.2019.

[103] Amanda Lenhart, Maeve Duggan, Andrew Perrin, Renee Stepler, Harrison Rainie, Kim Parker, et al. *Teens, social media & technology overview 2015*. Pew Research Center [Internet & American Life Project], 2015.

[104] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.

[105] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80. IEEE, 2012.

[106] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE, 2011.

[107] Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, João Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[108] Tech vs Abuse (Comic Relief). Tech vs abuse: Research findings, 2019.

[109] Gary A Klein. *Streetlights and shadows: Searching for the keys to adaptive decision making*. MIT Press, 2011.

[110] Paul Ekblom. Designing products against crime. *Encyclopedia of Criminology and Criminal Justice*, pages 948–957, 2014.

[111] Ronald Victor Gemuseus Clarke. *Situational crime prevention*. Criminal Justice Press Monsey, NY, 1997.

[112] John E Eck. *Assessing responses to problems: An introductory guide for police problem-solvers*. US Department of Justice, Office of Community Oriented Policing Services, 2004.

[113] SaveLives. For Maracs. Accessed: 08.18.2019.

[114] Mariana Gerber and Rossouw Von Solms. Management of risk in the information age. *Computers & Security*, 24(1):16–30, 2005.