Identifying Unintended Harms of Cybersecurity Countermeasures

Yi Ting Chua*, Simon Parkin†*, Matthew Edwards‡, Daniela Oliveira§,
Stefan Schiffner¶, Gareth Tyson‖ and Alice Hutchings*
*University of Cambridge, {*yiting.chua, alice.hutchings*}@cl.cam.ac.uk
†University College London, s.parkin@ucl.ac.uk
‡University of Bristol, matthew.john.edwards@bristol.ac.uk
§University of Florida, daniela@ece.ufl.edu
¶University of Luxembourg, stefan.schiffner@uni.lu
‖Queen Mary University of London, g.tyson@qmul.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 illuminate harmful consequences, not to paralyze decision-making, but so that potential unintended harms can be more thoroughly 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 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 up-to-date (e.g., to thwart commodity attacks), to targeted countermeasures to address risks specific to an individual system or situation of concern.

Countermeasures can include technical controls, such as advanced verification of user accounts. Countermeasures can also include policies and guidance for the users of a system, such as awareness materials or a declaration of expected conditions of use (as for a forum or an organisation's computers). 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. However, countermeasures themselves may introduce unintended consequences, be it in crime prevention [1]–[3], physical safety [4], or in IT-security more generally [5]. Often even less considered: in fact countermeasures can actually do *harm*, e.g. to infrastructure or to some or all of its users. This harm may be as slight as causing disruption and additional security burden for legitimate users of a system [6], or as severe as causing negative impact on 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 detail is demonstrated in recent real-world events. One example is the deployment of facial recognition in publicly-accessible spaces to augment law enforcement and public control capabilities. The intention may have been to reduce crime and unwanted behaviours, but it has also sparked privacy concerns, for instance in the United States [7] and United Kingdom [8]. The debate centres on whether such systems are appropriate, given the potential for invasion of privacy and linking of data to other systems (with, in some cases, limits to system accuracy [9], which could have potentially harmful consequences). This example shows where a risk countermeasure with protective aims 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 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). Stefan Schiffner is supported by the Luxembourg National Research Fund as part of EnCaViBS, a CORE project.
*Authors contributed equally.

A. Our contribution

Our contributions are arranged in three steps: case studies, a framework based on outcomes of the case studies, and application of the framework to a new case study. We begin by presenting five case studies (Section III) to help highlight readily observed unintended harms. We select a range of complex examples, namely: intimate partner abuse, disinformation campaigns, CEO fraud, phishing, and dating fraud. For each case study, we consider the unintended consequences associated with the potential interventions that may be applied. 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*.

Based on the case studies, we construct a framework for conceptualising potential unintended consequences and harms (Section IV). We classify these as: imposing additional cost; misuse; insecure norms/complacency; false positives; displacement; amplification; or disruption of other countermeasures. To evaluate our findings, 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 harms. 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 harms are monitored alongside the intended outcomes. Considering other approaches which may illicit unintended harms (Section VIII), we discuss the implications and potential future applications of the framework (Section VII), and conclude with next steps (Section IX).

II. Background

Here we outline how content (produced by user behaviour), users, and infrastructure may be impacted by countermeasures, and the unintended consequences of particular countermeasures (including unintended harms). Risk may be managed in a centralised manner (such as security in an organisation/business, often managed by a security manager or security function). In more complex situations, risk management efforts must involve a range of stakeholders (including end-users) across a wider environment.

A. Countermeasures

We consider a treatment specifically deployed to handle a risk countermeasure. Referring to the ISO 27001 risk treatment framework [10], action may be taken to manage a risk, or accept it as a residual risk. 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.

Fig. 1. The Johari Window, as a means to consider the limitations to knowledge of risks (and in turn, countermeasures) between one entity and others in the ecosystem - reproduced from [12].

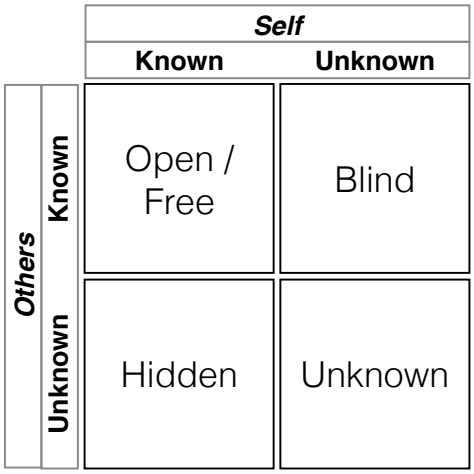

Figure 1 [12] depicts the different qualities of ignorance a stakeholder may suffer from as the Johari Window [13]. This is where *unintended consequences* can emerge: if action is taken based on incomplete knowledge, it may increase consequences for a particular group which was not the intended target, to the extent that it creates *harms*. A risk/system owner may be best-served by gathering information or opinions from other stakeholders in the environment before taking action. This can increase the open/free knowledge available to many, as in Figure 1.

B. Unintended consequences

Unintended consequences can refer to observed phenonema such as 'knock-on effects', 'side-effects', or 'maladaptive responses' (e.g., to security awareness campaigns [14]). A countermeasure may also trigger a 'cascade effect' [15], in what appears to be a sequence of subsequent actions. It is then necessary to have a means to map this complexity.

Relevant principles considered in the economics of security can 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, for instance, 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, including those who were not the intended target. 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 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 ongoing risk measurement capabilities (referring to Figure 1, the risk owner becomes 'blind').

Another principle considered in security economics is *information asymmetry* [18]. That is where information or actions are known by one entity but not others. Those impacted by risk management actions may not have the information, which risk owners might assume they would have, to manage risks appropriately. Such 'hidden' actions can also 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 a well-meaning system owner who acts upon the asset (such as disrupting a website, 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 the avoidance of harms becomes part of the strategic planning of the platform itself. One example may be if controls prevent some users from using an online social platform, even where the impact on a subset of those users is in effect 'collateral damage'— the controls have achieved their intended aim, but also impacted others who it was not intended to affect.

C. Unintended harms

We regard *unintended harms* as unintended consequences which 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 burdened with as a result. 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 prior examination of risk management has alluded to potential side-effects of cybersecurity countermeasures, we believe this is the first work to consider development of a dedicated framework for (i) exploring potential *unintended harms*, and (ii) prioritising their identification to protect users from being adversely (perhaps irrecoverably) affected.

III. Case Studies

We describe five *case studies* within the space of 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) may hamper any efforts by Bob to access housing and financial support [28]. If tech-enabled communication were not disrupted, 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), legal channels may be developed 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, to work on behalf of Bob [27].

Provide guidance. It may seem useful for experts to *produce advice* for securing personal devices, so that people in a similar situation to Bob can control access to shared devices and accounts. 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 (more so if those devices are shared [29]). 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 not be an appropriate way to provide support.

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 main goal of the campaign is to deceive voters [30]–[32].

Content removal. This countermeasure generally involves the removal of content, accounts and/or bots [30], [31]. The removal of content may create a '*Streisand effect*', where the request to remove content can draw increased attention to it [30], [33]. In the scenario, removal of content may *backfire* if the third-party presents this as unjust [33], using it as proof of a conspiracy and suppression of 'truth' against them. Such removal can potentially speed up misinformation diffusion [34], [35]. Removing only some misinformation while leaving similar material available potentially confers validity to that remaining content [36], making it more difficult for citizens to determine which content can be trusted.

Account removal. The removal of accounts or content does not address the root cause or motivation for a misinformation campaign. Instead, it *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 on mainstream platforms such as Facebook and Twitter [37]. This may in turn result in an 'echo chamber', where individuals are surrounded with information that confirms their own beliefs, opinions, and views, and ultimately results in group polarisation [32], [38].

Removal of bots. Although potentially effective [39], this can result in *misclassification*. Misclassification is of increasing concern as social bots' capabilities to generate human-like behaviours are improving [40], [41]. There are two general types of misclassification: false negatives and false positives. *False negatives*, or the misclassification of bots as legitimate accounts, can intensify the effects of disinformation, as users may trust information from bots [42], and bots have been found to be more likely to share false information [39]. *False positives*, or misclassification of non-bot accounts as bots, can lead to a perception of censorship among legitimate users [39]. It can also potentially displace users to other platforms (then transferring risks, rather than reducing them).

Automated detection algorithms. The development of auto-detection algorithms [31], [39], [40] 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 [30], [39]. This can, perversely, potentially reduce users' scepticism towards misinformation [42]. 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 [41].

Fact-checking. Fact-checking [32] may be introduced. This may either incorporate fact-checking as part of content management [31], or encourage users to utilise tools prior to sharing information [42]. 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 [43], and by content of the fact-check [44]. In addition, the effectiveness of fact-check posts are dependent on the content's level of controversy [44]. 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 [43]. 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 their actions.

Behaviour and security culture change. Changes may be made to working practices, to remove the likelihood of employees finding themselves under pressure to fulfil last-minute requests that do not follow correct protocols. This can lead to *complacency*, if restrictions through technical solutions are not matched with changes to workplace behaviours (foremost, senior staff making last-minute requests from non-corporate email accounts). There are a range of challenges in measuring security culture [45], 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 can 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 request or execute transfers without checks. For instance, additional authentication may be requested 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 each 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's supplier contacts 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 company's 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 URLs) to facilitate detection of phishing messages [46], [47]. Internet Service Providers (ISPs) or email providers block the detected messages from being delivered. System users may become *complacent* about the emails that they have sight of, assuming that automated solutions block anything that is potentially harmful to them (although automated solutions do not necessarily block all malicious emails).
Website takedown. Websites used for phishing may also be taken down [48]. 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. Conversely, *false positives* resulting from message filtering 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 and education. Attempts may be made to educate users so that they can identify phishing messages [49]–[51]. There are a variety of proposed training approaches, ranging from games [52], [53] to simulated phishing campaigns, especially in corporate environments [50]. It can be the case that those who are unable to identify phishing messages repeatedly fail to spot them [50], masking that alternative solutions are needed. Training can also induce *additional costs* to users, in attending training sessions and then managing additional security tasks on top of their primary work tasks everyday activities. This adds to the users' 'compliance budget' [54], reducing productivity. For certain users, it might also cause them to become overly sensitive, ignoring legitimate messages they are unsure how to classify (where this can increase *false positives*). Another general unintended harm of user training is that *attackers may adapt to act around prescribed behaviours*, in something of 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 [55], and crucially, legitimate users may find that the verification process *interferes with their preferred means of self-representation* on a dating site, which itself could involve some small degree of misrepresentation or selectivity [56]. 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 [57]), and correspondingly increase *safeguarding responsibilities* for these organisations [58].
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 [55], [59]. Reporting systems can be *abused* by users in redress of personal grievances [60], and so reports must be reviewed by human moderators, a monotonous job which may create additional risks and harms [61]. 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) [55], 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 in some cases expose the victim to additional risk of *revictimisation* fraud [55].
Provide advice. Quite aside from whether the advice for avoiding fraud is effective [62], well-intentioned descriptions of 'what to look out for' can provide invaluable advice for fraudsters on how to *disguise their future activities on dating platforms*.

F. Summary

Our case studies have illustrated an – albeit limited – range of unintended harms emerging from otherwise well-meaning countermeasures to risks. In exploring unintended harms across the case studies, we recognise overarching categories of unintended harms across multiple scenarios (e.g.,

that automated solutions may misclassify legitimate users and their activity, or that increased security may disrupt how well-meaning users interact with an online platform). These categories were developed into a taxonomy that can support coordinated exploration of 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 taxonomy of unintended harms of cybersecurity countermeasures. From this, we discuss a simple framework which can be used 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 and points to their potential harms. 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 [63]. Examples include the surge of new online drug markets following the takedown of Silk Road [64], or phishing sites moving to domains and hosts which are more resistant to takedown efforts [48].
2) **Insecure norms**: The implementation of countermeasures encourages insecure behaviours, creating the potential for greater harm. Examples include creating a reliance on technical controls [65], 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 [66], the costs to some stakeholders may outweigh the original harm. Examples include reporting systems for social media abuse which pose a burden of manual review for social media companies and their employees [61], and cases where there is a reliance on anti-phishing training which places the burden of responsibility for phish detection 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 [67]. Examples include reporting systems being used maliciously as a result of personal grievances [60] or competitive business interests [48], 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 it is not anticipated. Examples include the "cold start" problem in reputation systems, with legitimate new users being unable to establish credibility to enter a community, 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 [68], and the 'Streisand effect', where an attempt to take down content causes increased interest in preserving and sharing it [33].
7) **Disruption**: 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 confusion and mistakes among users [69], [70].

B. A Framework for Unintended Harms

We further developed the unintended harm categories described above into a framework of questions. These questions may be asked of any (proposed or existing) countermeasure, in order to identify potential negative and harmful 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. This is especially important when a countermeasure might 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 be open enough to prompt consideration of any or all of *behaviours*, *users*, and *infrastructure*, beyond thinking of implications for technological solutions alone. We do not assert that the framework will exhaustively identify all dimensions of unintended consequences equally, but that the questions in the framework sufficiently support exploration of impacts which relocate or transform user behaviours, users and their circumstances, and infrastructure (including how they are perceived). One goal is to provide decision-makers with the tools to anticipate the (potentially harmful) unintended consequences of their actions, where currently there are few, if any, tools which achieve this directly.

The framework is intentionally *generative*, providing prompts for the identification of new possible harms from

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?

a countermeasure. This does not prescribe 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. The framework could potentially be employed alongside existing risk assessment frameworks, such as ISO27001 or the NIST risk management guidelines [71].

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 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 in the service or environment can then be addressed by looking at the process which produces 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 [72]. 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. For some countermeasures, the design and implementation may inevitably lead to some degree of harm. However, where such harms are unintended, and thus unexpected and unknown, they may be excluded from a risk assessment activity and not managed appropriately. This can lead to decisions being taken in an under-informed manner (referring to the 'Unknowns' in Figure 1, pointing to opportunities to become more informed). Our framework aims to generate these consequences, not to paralyze decision-making, but so that consequences can be more thoroughly anticipated in risk management strategies.

V. VULNERABLE POPULATIONS

A countermeasure may be deployed to manage a risk, with this realising a benefit for (remaining) users of the managed service/system. Considering the different unintended harms described in Section IV, not all users will experience the benefits of a countermeasure in the same way. We may see that some discernible user groups are protected by countermeasures and experience their benefits, while other groups are left not served by the countermeasure, lacking support, or at worst harmed directly by the countermeasure. This is unintended harm if the risk owner does not know that they have disadvantaged or negatively affected user groups which they did not knowingly intend to control.

Referring to Table II, unintended harms can include impacts to User interactions with a Service/System (their observable behaviours), their access (whether and how they are regarded as Users in the first instance), and changes to the infrastructure (such as whether particular Service controls or Shared controls are present or remain in place). The absence of an assumed control, user group, or behaviours can also result in harms (for example, that users on the same platform will respect each other's privacy, and that the Service and risk owner will ensure this).

We are then considering 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. It would also be assumed that they would 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 — if adverse effects are discovered after deployment of a control, damage may have already been done.

TABLE II
THE ELEMENTS OF AN OBSERVED SYSTEM AND HOW THEY INTERACT.

Element	Description	Element Type
Ecosystem	A collection of **service**, **user** and **user group** instances	ALL
Service	Managed service (or system), including assets	Actor
User	Individual user, who may or may not be in a **user group** or be using a **service**	Actor
User Group	Group of **users** sharing attached characteristics	Actor
User+User Actions	Users may interact with each other within or outside a **service**, individually or within a **user group**	Behaviour
User+Service Actions	Behaviour exhibited by a **user** when interacting with a **service** (including the data produced as a result)	Behaviour
Service control	A control that exists only within the domain of a **service**	Infrastructure
Shared control	A control which exists within the **ecosystem**, and may be applied to support a **service** (un)knowingly	Infrastructure

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

(i) May not have the access or capabilities needed to make use of provisioned controls, thereby missing the intended benefit and continuing to live with the risk. They may need access to more appropriate, *alternative controls* (relating to outcome (a), Section IV).
(ii) May be affected by a combination of harms, which may all affect a user or user group at once. The more distinct harms which affect a group, the more the group should be considered as needing concerted assistance. This relates to outcome (b), Section IV.
(iii) May be 'forgotten', especially if they are (inadvertently) removed from the service (or ecosystem) or their intended behaviours are restricted. Design accommodations could then be considered in advance. This relates to the masking of problems (outcome (c), Section IV), to the extent where representation/advocacy outside of technology may be necessary.

Referring to Table II, for (i) 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 specifically, 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 has been removed). An appropriate control for impacted users may be unavailable, unreachable, or may not yet exist.

For (ii) above, any combination of harms (as in Figure II) may impact a user or user group. This could, for instance, be a combination of usability problems in user-facing technologies, a lack of preparedness and skills support for users, and the malicious intentions of a determined (i.e., skilled) attacker/perpetrator with access to the same technologies.

For (iii) 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. If user data is created and retained in the service environment, the *absence or lack* of user-related data or user accounts is then of increased relevance. This then requires 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 a company rely informally on IT support at their workplace, which they no longer have access to when they retire [73].

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 [74]. This is referred to as the 'differential effects' of information security countermeasures. We explore the differential effects upon users who the infrastructure owners would *not* want to be in *a state which has the same or more exposure to risk or a lack of security as it had before deployment of the countermeasure*.

Here we consider a range of vulnerable user groups, which is not intended to be exhaustive, but instead demonstrates 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. For these reasons, harms to distinct populations can include *risk dumping* upon users who are not prepared or supported; an *unpredictable/destabilised cyber-physical environment* which creates new risks, or; *masking of risks* from the view of the risk owner, though the risks persist for the user.

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 (to which they may be regarded as legitimate users like any other). Nonetheless, these users could have been spared the additional cost if the events leading to their initial exclusion were preempted and avoided. Referring to concepts represented across Figure 1 and Table II, 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.
Vulnerable populations (such as in our Case Studies, Section III) can include:

- **Older adults.** May have had fewer opportunities to habituate use of technologies compared to younger groups who grew up with computing technologies. Older adults may however also have a heightened sense of risks. These users may be 'hidden' if they interact with technology support rarely – for instance, visiting a physical retail store to buy a new device relatively rarely [73], delegating device maintenance to a paid 'IT person', or rarely interacting with (positive/enabling) technologies [75].
- **Small businesses.** Smaller businesses may have fewer resources available to invest in automated security solutions, and be less likely to have a dedicated security function to manage threats [76]. For small charities also, this can include relying on volunteers having the necessary expertise [77]. Smaller organisations may also delegate security to an expert IT-security provider company [77].
- **Survivors/victims of tech-abuse.** May be controlled or monitored both physically and through (forcibly shared) devices and online services. Opportunities to configure or modify devices may be limited, and there are potential implications if a perpetrator discovers device activity which they believe interferes with their control [29]. There are frontline services skilled in addressing domestic and intimate partner abuse, where otherwise those suffering tech-abuse may be 'hidden' from view; the configuration of shared consumer devices is often assumed to be agreed between all members of a shared living space.
- **People with disabilities or impairments.** Security and privacy controls are often developed for the general population [78]. This can inadvertently sideline not only groups such as those highlighted above, but also users with disabilities (such as visual impairments). This highlights the need to manage risks while also ensuring universal access to services, so that all intended users can benefit.

VI. Framework Application: 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 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 [79].

Various stakeholders have developed and implemented a range of cyberbullying countermeasures (e.g., [19], [80], [81]). Countermeasures in this space are not necessarily risk-free, and merit careful assessment given that one of the target audience groups is young users [82], [83].

In some instances, the unintended harms of the countermeasures outweigh the benefits. For example, parents suggesting children to include false information such as age with accurate information in online profiles can interfere with automated detection and filtering systems [84]. Thinking to the long-term effects of such strategies, these countermeasures have the potential to change behaviours and how younger users interact with technology. They then must be considered carefully.

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 [85]–[87]. Cyberbullying victimisation is shown to correlate with an array of negative consequences. For example, Hinduja and Patchin [88] found that individuals who were cyberbullied were more likely to report offline problem behaviours such as running away from home or carrying a weapon. Females were also more likely to be victims of cyberbullying [86], [89]. 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 [90].

To limit these negative outcomes, there are at present a multitude of countermeasures. To illustrate the applicability of the proposed framework, we will focus on two common countermeasures for cyberbullying – education and training (Section VI-A), and privacy control and management (Section VI-B). For each countermeasure, we identify potential unintended harms.

A. Education and training — unintended harms

Education and training programs are frequently recommended to various stakeholders, such as teenagers, parents, and educators [19], [80], [81], [83]. For teenagers, the goals of the programs, which tend to be administered school-wide, are to establish basic knowledge on cyberbullying and appropriate online behaviours, and communicate the consequences of cyberbullying without distinguishing the bullies from the bullied [83], [91], [92]. One exception is Facebook's factsheet where such distinction was taken into consideration [80].

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 [91], [92]. There are also programs that place an emphasis on protective factors, such as a positive school climate [93] and resilience [94], 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 [80], [81]. 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 [87], [95].

- **Insecure norms.** Education and training 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 [83], [92], [96]. For instance, most victims of cyberbullying do not disclose to adults [86] or utilise the block function of online communication tools [97]. 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.** There is an extra burden on stakeholders to develop and implement the countermeasure, in terms of effort, resources and time. Teachers and educators need to allocate time to attend training sessions and/or become trainers for other staff in schools [19], [98]. For school administrators, the burden lies in coordinating and incorporating these programs into existing curriculum and community involvement [19], [91]. Recommended practices often emphasise the role of schools in initiating education and training programs [19], [81]. In providing knowledge to users of technologies, one challenge is to ensure that education and training needs to keep pace with technologies as they emerge, but also the ways in which technologies are used.
- **Misuse.** The knowledge and information made available through education and training, especially school-wide programs, may potentially be used by perpetrators. Engagement in cyberbullying behaviours may correlate with having been a victim of cyberbullying previously [79], [86]. Individuals who attend these programs would now have knowledge on techniques for cyberbullying.
- **Misclassification.** Incorrect classification arises when definitions of cyberbullying (which lack consensus [83]) 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 [99].
- **Amplification.** This countermeasure 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 [100]. To illustrate, consider a form of direct bullying where the cyberbully sends an email with a malicious attachment [79]. The recipient who opens the email may be blamed, as education programs specifically warn individuals not to open suspicious emails [82]. 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 prescribed use of technology.
- **Disrupting other countermeasures.** With a multi-stakeholder approach on education and training [19], [81], [83], the likelihood of confusion and contradictory information is high. The current lack of consensus on the definition of cyberbullying [83] means that students may potentially be receiving different advice around cyberbullying behaviours from their parents, teachers, and social media platforms.
- **Vulnerable population.** 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 [100]. The second group includes pre-teens and adolescents who experience physical isolation and rely on online communities for social support. Arguably, young people ought to be able to have a positive security experience in using technologies, and be able to maintain healthy interactions with others without security precautions becoming overly restrictive. For example, adolescents diagnosed with cancer may use online forums to share experiences and cope with emotions [101].

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 settings and controls for devices, applications, and social media platforms [80], [81].

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 [92]. Parents have also suggested to their children to blend false information, such as age, when sharing personal information in online profiles as a privacy management technique [84], or rely on applications that promote online safety via monitoring [102].

Applying the framework, some potential unintended harms of *privacy control and management* are as follows:

- **Displacement.** There are two potential types of displacement. First, the countermeasure may encourage migration to other types of platforms with more approachable privacy control settings (pre-teens have reported a lack of awareness or understanding of privacy settings on sites such as Facebook [84]). An example is the migration to Snapchat [103], which advertises straightforward privacy features such as deletion of content [104]. Second, this countermeasure may displace harms to individuals without proper privacy settings and controls, or who are less able to use them.

- **Insecure norms** 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 [104]. For other applications such as Facebook, users may be relying on the default privacy settings to protect them [84].
- **Additional costs.** Privacy control and management brings additional costs and effort to pre-teens and teenagers. They need to dedicate time for learning about privacy setting controls across devices, applications and platforms [80], [81]. 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 [105]. This means that they may either be relying on the default settings, or not take the time to learn how to increase account privacy.
- **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 [79]. This can overwhelm individuals as they monitor their online presence. Second, cyberbullies can isolate targeted individuals by requesting the removal of these individuals' legitimate information and accounts that can be found via search engines.
- **Misclassification.** One scenario where there may be unintended consequences is parental privacy control and management [92]. 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 [33]. 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 [84], may interfere with countermeasures that rely on automated detection and filtering [96], [106]–[109]. 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 that they engage in safe practices, such as adjusting privacy settings on social media sites, but at the same time, there are individuals who seem to be unaware of such settings [84].

VII. Discussion

Application of the framework can support identification of appropriate interventions when signs of risks emerge in a socio-technical system, such as when online relationship activity begins to include unhealthy behaviours [110].

Looking specifically at service and technology design, it can be possible to feed the outcomes of applying the framework into efforts to, for instance, Design Against Crime [111]. 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 [112] and problem-oriented policing [113] 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 to identifying harm-free interventions 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) [114] 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 [115]. With the proposed framework, an assessment of residual risks will also include examining risks which emerge because of the deployment of choices in 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 reasoned consideration of the side effects with the benefits. Within this is a need to consider decision-maker *preferences*, where harms should not be induced upon user groups who are considered to be unprepared or unsupported (Section V).

VIII. Related Work

Research into *personas* in information security [116], [117] encourages the representation of users by building a narrative of their perspective on a system. When personas are built

though user engagement, they can identify user previously unknown user requirements in a cyberphysical system [116], where our framework directly challenges system owners to inspect where such requirements may be impacted by risk controls themselves. Personas used in risk analysis may be built on the implicit assumptions of system owners [117], where these assumptions may need to be challenged through discourse with other stakeholders.

Similarly, *use cases* and *misuse cases* can be explored together to identify security requirements [118]. This process may be repeated as assets, risks, and controls are identified within an environment of users and the services they use.

Threat modelling may be used to increase situation awareness (for instance, with large-scale cyber command teams [119]), prompting practitioners to consider critical service capabilities and related threats. Our framework prompts stakeholders to consider critical users and user behaviours alongside capabilities which require protection, to complement existing efforts to manage risks to capabilities and assets. Premortems [120] encourage consideration of actions which can be taken in the present, to avoid existential risks to services in the future. Our framework, similarly, prompts consideration of what action could be taken in advance to insulate a process (for instance, a service or technology) against failure, in this case weaving unintended harms into the discussion.

From a different perspective, direct studies of stakeholders can also yield information about risks and unintended consequences; this can include surveys, interviews, and focus groups. Engagement with cyber security experts [121] has, for instance, identified expectations for top user behaviours (highlighting also that even the experts cannot necessarily agree on how users should be supported and protected). Interviews with organisational security managers [122] has identified that they can appreciate the link between security controls and impacts upon user tasks and priorities; in other cases, it has also identified that security managers themselves need tools to be able to consider the impacts that their risk management strategies have upon users [123].

Most current risk assessment and management approaches place emphasis on the identification and management of risks and threats without recognition of the potential harms such efforts may impose on user behaviours, users and infrastructure once countermeasures are deployed. Quantitative modelling [124] and cost-benefit analysis [71], [125], [126] approaches primarily consider how to directly address malicious threats. For example, the AEGIS framework considers countermeasures against the likelihood of attacks and cost-benefit analysis [126], alluding to side effects of countermeasures as costs.

The Cynefin framework [127] prompts risk owners to consider the complexities of a managed system and identify appropriate responses to cybersecurity risks. The authors emphasise that where cause and effect are not immediately apparent, there can be a need to consult with and consider the perspectives of others within the system. The framework for reducing human-related risks developed by Islam et al. [128] is motivated in part by the increased connectedness of individuals and technologies. The authors also approach cybercrime in terms of mobilising crime preventers within the system. Our unintended harms framework complements this goal by prompting risk owners to consider where the needs of 'good' system actors must be respected in advance of selecting countermeasures.

Many approaches to identifying or managing risks do not explicitly act to identify unintended consequences, let alone unintended harms. The proposed framework on unintended harms aim to equip stakeholders with a systematic approach to directly relate the consideration and assessment of unintended impacts — to service/system owners, the service itself, and service users in a complex ecosystem of increasingly interconnected services and user populations — when developing and deploying cybersecurity countermeasures.

IX. Conclusions

Here we have identified and charactertised the potential unintended harms of various cybersecurity risk countermeasures. Unintended harms are of increasing importance, firstly, as they can potentially have far-reaching impacts in current society where technology and the Internet are highly incorporated into our daily lifestyles. A countermeasure deployed to secure one aspect of interaction with a cyber-physical system may adversely affect other aspects of that same system. Countermeasures may prohibit or move behaviours, users, and elements of infrastructure completely from view or to another place in the larger ecosystem. We have shown how it is important to deliberately and proactively explore the potential for unintended harms.

Second, as illustrated in our case studies, the deployment of countermeasures related to cybersecurity and cybercrime often involves multiple agencies and stakeholders. The need for coordinated efforts from stakeholders adds another layer of complexity when assessing unintended harms and consequences. This is especially the case if a lack of coordination and communication between stakeholders can undo the risk management efforts of others.

There is a need for a strategic approach to uncover cyber-physical and socio-technical implications of any one cybersecurity risk intervention. We have illustrated the capacity of the framework in considering the unintended harms of candidate countermeasures, as well as its potential application as part of broader risk management strategies. Future work will apply the framework toward limiting unintended harms to society and its constituent user groups (including vulnerable populations).

Acknowledgements

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] P. N. Grabosky, "Unintended consequences of crime prevention," in *Crime Prevention Studies*. Citeseer, 1996.

[2] J. McCord, "Cures that harm: Unanticipated outcomes of crime prevention programs," *The Annals of the American Academy of Political and Social Science*, vol. 587, no. 1, pp. 16–30, 2003.

[3] B. C. Welsh and D. P. Farrington, "Toward an evidence-based approach to preventing crime," *The Annals of the American Academy of Political and Social Science*, vol. 578, no. 1, pp. 158–173, 2001.

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

[5] S. Pfleeger and R. Cunningham, "Why measuring security is hard," *IEEE Security & Privacy*, vol. 8, no. 4, pp. 46–54, 2010.

[6] A. Adams and M. A. Sasse, "Users are not the enemy," *Communications of the ACM*, vol. 42, no. 12, pp. 41–46, 1999.

[7] D. Lee, "San Francisco is first us city to ban facial recognition," BBC, 2019, accessed: 08.15.2019. [Online]. Available: https://www.bbc.co.uk/news/technology-48276660

[8] Z. Kleinman, "Facial recognition in King's Cross prompts call for new laws," BBC, 2019, accessed: 08.15.2019. [Online]. Available: https://www.bbc.co.uk/news/technology-49333352

[9] BBC News, "2,000 wrongly matched with possible criminals at champions league," BBC, 2018, accessed: 08.16.2019. [Online]. Available: https://www.bbc.co.uk/news/uk-wales-south-west-wales-44007872

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

[11] J. Adams, *Risk*. University College London Press, 1995.

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

[13] J. Luft and H. Ingham, "The Johari window," *Human relations training news*, vol. 5, no. 1, pp. 6–7, 1961.

[14] G. Stewart and D. Lacey, "Death by a thousand facts: Criticising the technocratic approach to information security awareness," *Information Management & Computer Security*, vol. 20, no. 1, pp. 29–38, 2012.

[15] M. G. Porcedda and D. S. Wall, "Cascade and chain effects in big data cybercrime: Lessons from the talktalk hack," in *Proceedings of WACCO 2019: 1st Workshop on Attackers and Cyber-Crime Operations, Held Jointly with IEEE EuroS&P*, 2019.

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

[17] I. Kirlappos, S. Parkin, and M. Sasse, "Learning from "shadow security": Why understanding non-compliant behaviors provides the basis for effective security," in *USEC '14 Workshop on Usable Security*, 2014, pp. 1–10.

[18] R. Anderson and T. Moore, "Information security economics–and beyond," in *Annual International Cryptology Conference*. Springer, 2007, pp. 68–91.

[19] M. A. Couvillon and V. Ilieva, "Recommended practices: A review of schoolwide preventative programs and strategies on cyberbullying," *Preventing School Failure: Alternative Education for Children and Youth*, vol. 55, no. 2, pp. 96–101, 2011.

[20] I. Lopez-Neira, T. Patel, S. Parkin, G. Danezis, and L. Tanczer, "'Internet of Things': How abuse is getting smarter," *Safe–The Domestic Abuse Quarterly*, vol. 63, pp. 22–26, 2019.

[21] L. A. Gordon, M. P. Loeb, and T. Sohail, "A framework for using insurance for cyber-risk management," *Communications of the ACM*, vol. 46, no. 3, pp. 81–85, 2003.

[22] M. Levi, A. Doig, R. Gundur, D. Wall, and M. Williams, "Cyberfraud and the implications for effective risk-based responses: themes from UK research," *Crime, Law and Social Change*, vol. 67, no. 1, pp. 77–96, 2017.

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

[24] D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. Dell, ""A stalker's paradise": How intimate partner abusers exploit technology," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 667.

[25] N. Sambasivan, A. Batool, N. Ahmed, T. Matthews, K. Thomas, L. S. Gaytán-Lugo, D. Nemer, E. Bursztein, E. Churchill, and S. 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*. ACM, 2019, p. 2.

[26] M. Emms, B. Arief, and A. van Moorsel, "Electronic footprints in the sand: Technologies for assisting domestic violence survivors," in *Annual Privacy Forum*. Springer, 2012, pp. 203–214.

[27] D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. 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*, vol. 1, 2017.

[28] Tech vs Abuse (Comic Relief), "Tech vs abuse: Design principles," Tech vs Abuse, 2019, accessed: 07.28.2019. [Online]. Available: https://www.techvsabuse.info/design-principles

[29] S. Parkin, T. Patel, I. Lopez-Neira, and L. Tanczer, "Usability analysis of shared device ecosystem security: Informing support for survivors of IoT-facilitated tech-abuse," in *Proceedings of the New Security Paradigms Workshop (NSPW '19)*. ACM, 2019.

[30] Twitter Inc., "Retrospective Review: Twitter, Inc. and the 2018 Midterm Elections in the United States," Twitter,Inc., Tech. Rep., February 2019, accessed: 08.12.2019. [Online]. Available: https://blog.twitter.com/content/dam/blog-twitter/official/en_us/company/2019/2018-retrospective-review.pdf

[31] Facebook Newsroom, "What is Facebook doing to address the challenges it faces?" Facebook Newsroom, February 2019, accessed: 07.28.2019. [Online]. Available: https://newsroom.fb.com/news/2019/02/addressing-challenges/

[32] G. Haciyakupoglu, J. Hui, V. S. Suguna, D. Leong, and M. F. Bin Abdul Rahman, "Countering fake news: A survey of recent global initiatives," Nanyang Technological University, Tech. Rep., 3 2018.

[33] S. C. Jansen and B. Martin, "The Streisand effect and censorship backfire," *International Journal of Communication*, vol. 9, pp. 656–671, 2018.

[34] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[35] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[36] T. Caulfield, J. Spring, and A. Sasse, "Why Jenny can't figure out which of these messages is a covert misinformation operation," in *Proceedings of the New Security Paradigms Workshop (NSPW '19)*. ACM, 2019.

[37] E. Livni, "Twitter, Facebook, and Insta bans send the alt-right to Gab and Telegram," Quartz News, May 2019, accessed: 05.12.2019. [Online]. Available: https://qz.com/1617824/twitter-facebook-bans-send-alt-right-to-gab-and-telegram/

[38] C. R. Sunstein, *Republic.com 2.0*. Princeton, NJ: Princeton University Press, 2007.

[39] C. Shao, G. Luca Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature Communications*, vol. 9, no. 4787, pp. 1–9, 07 2018.

[40] E. Ferrara, O. Varol, C. David, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 07 2016.

[41] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 06 2010.

[42] J. M. Burkhardt, *Combating Fake News in the Digital Age*, ser. 8. American Library Association, 2017, vol. 53.

[43] A. Hannak, D. Margolin, B. Keegan, and I. 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*, pp. 187–196, 01 2014.

[44] R. K. Garrett, E. C. Nisbet, and E. K. Lynch, "Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naive theory," *Journal of Communication*, vol. 63, no. 4, pp. 617–637, 2013.

[45] ENISA, "Cybersecurity culture guidelines: Behavioural aspects of cybersecurity," 2019, accessed: 08.15.2019. [Online]. Available: https://www.enisa.europa.eu/publications/cybersecurity-culture-guidelines-behavioural-aspects-of-cybersecurity

[46] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A Survey of Phishing Email Filtering Techniques," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 2070–2090, 2013.

[47] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious urls: An application of large-scale online learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 681–688.

[48] A. Hutchings, R. Clayton, and R. Anderson, "Taking down websites to prevent crime," in *2016 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2016, pp. 1–10.

[49] P. Kumaraguru, Y. Rhee, S. Sheng, S. Hasan, A. Acquisti, L. F. Cranor, and J. 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*, ser. eCrime '07. New York, NY, USA: ACM, 2007, pp. 70–81. [Online]. Available: http://doi.acm.org/10.1145/1299015.1299022

[50] D. Caputo, S. Lawrence Pfleeger, J. D. Freeman, and M. Johnson, "Going spear phishing: Exploring embedded training and awareness," *Security & Privacy, IEEE*, vol. 12, pp. 28–38, 01 2014.

[51] O. A. Zielinska, R. Tembe, K. W. Hong, X. Ge, E. Murphy-Hill, and C. 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*, vol. 58, no. 1, pp. 1466–1470, 2014.

[52] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and E. 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)*, vol. 229, 01 2007, pp. 88–99.

[53] N. Arachchilage and S. Love, "A game design framework for avoiding phishing attacks," *Computers in Human Behavior*, vol. 29, p. 706714, 05 2013.

[54] A. Beautement, M. A. Sasse, and M. Wonham, "The compliance budget: managing security behaviour in organisations," in *Proceedings of the 2008 New Security Paradigms Workshop*. ACM, 2009, pp. 47–58.

[55] A. Rege, "What's love got to do with it? exploring online dating scams and identity fraud." *International Journal of Cyber Criminology*, vol. 3, no. 2, 2009.

[56] N. Ellison, R. Heino, and J. Gibbs, "Managing impressions online: Self-presentation processes in the online dating environment," *Journal of computer-mediated communication*, vol. 11, no. 2, pp. 415–441, 2006.

[57] J. Birnholtz, C. Fitzpatrick, M. Handel, and J. 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*. ACM, 2014, pp. 3–12.

[58] K. Albury, J. Burgess, B. Light, K. Race, and R. Wilken, "Data cultures of mobile dating and hook-up apps: Emerging issues for critical social science research," *Big Data & Society*, vol. 4, no. 2, pp. 1–11, 2017.

[59] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically dismantling online dating fraud," *IEEE Transactions on Information Forensics and Security*, 2019.

[60] J. Anderson, M. Stender, S. M. West, and J. C. York, "Unfriending censorship," Onlinecensorship.org, Tech. Rep., 2016.

[61] S. T. Roberts, "Behind the screen: The hidden digital labor of commercial content moderation," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2014.

[62] M. T. Whitty, "Who can spot an online romance scam?" *Journal of Financial Crime*, vol. 26, no. 2, pp. 623–633, 2019.

[63] 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.

[64] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *24th USENIX Security Symposium (USENIX Security 15)*, 2015, pp. 33–48.

[65] A. Mylonas, A. Kastania, and D. Gritzalis, "Delegate the smartphone user? security awareness in smartphone platforms," *Computers & Security*, vol. 34, pp. 47–66, 2013.

[66] R. Bojanc and B. Jerman-Blažič, "An economic modelling approach to information security risk management," *International Journal of Information Management*, vol. 28, no. 5, pp. 413–422, 2008.

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

[68] C. Southworth, S. Dawson, C. Fraser, and S. Tucker, "A high-tech twist on abuse: Technology, intimate partner stalking, and advocacy," *Violence Against Women Online Resources*, 2005.

[69] H. Murray and D. Malone, "Evaluating password advice," in *2017 28th Irish Signals and Systems Conference (ISSC)*. IEEE, 2017, pp. 1–6.

[70] S. Tye-Williams and K. J. Krone, "Identifying and re-imagining the paradox of workplace bullying advice," *Journal of Applied Communication Research*, vol. 45, no. 2, pp. 218–235, 2017.

[71] G. Stoneburner, A. Y. Goguen, and A. Feringa, "NIST SP 800-30 risk management guide for information technology systems," 2002.

[72] D. Maimon and E. R. Louderback, "Cyber-dependent crimes: an interdisciplinary review," *Annual Review of Criminology*, vol. 2, pp. 191–216, 2019.

[73] S. Parkin, E. M. Redmiles, L. Coventry, and M. A. 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.

[74] J. D'Arcy and A. Hovav, "Does one size fit all? examining the differential effects of is security countermeasures," *Journal of Business Ethics*, vol. 89, no. 1, p. 59, Aug 2008.

[75] A. Frik, L. Nurgalieva, J. Bernd, J. Lee, F. Schaub, and S. 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: USENIX Association, Aug. 2019.

[76] S. Parkin, A. Fielder, and A. 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*. ACM, 2016, pp. 69–80.

[77] E. Osborn and A. Simpson, "Risk and the small-scale cyber security decision making dialoguea uk case study," *The Computer Journal*, vol. 61, no. 4, pp. 472–495, 2018.

[78] Y. Wang, "Inclusive security and privacy," *IEEE Security & Privacy*, vol. 16, no. 4, pp. 82–87, 2018.

[79] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims," *New Media & Society*, vol. 11, no. 8, pp. 1349–1371, 2009.

[80] Facebook, "Empower teens," 2016, accessed: 08.12.2019.

[81] S. Hinduja and J. Patchin, *Cyberbullying: Identification, Prevention, & Response*. Cyberbullying Research Center, 2018.

[82] S. Hinduja and J. W. Patchin, "Preventing cyberbullying: Top ten tips for teens," Cyberbullying Research Center, June 2018, accessed: 08.12.2019. [Online]. Available: https://cyberbullying.org/Top-Ten-Tips-Teens-Prevention.pdf

[83] R. A. Sabella, J. W. Patchin, and S. Hinduja, "Cyberbullying myths and realities," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2703–2711, 2013.

[84] K. Davis and C. James, "Tweens' conceptions of privacy online: implications for educators," *Learning, Media and Technology*, vol. 38, no. 1, pp. 4–25, 2013.

[85] J. W. Patchin and S. Hinduja, "Bullies move beyond the schoolyard: A preliminary look at cyberbullying," *Youth Violence and Juvenile Justice*, vol. 4, no. 2, pp. 148–169, 2006.

[86] Q. Li, "New bottle but old wine: A research of cyberbullying in schools," *Computers in Human Behavior*, vol. 23, no. 4, pp. 1777–1791, 2007.

[87] T. Lorenz, "Teens are being bullied 'Constantly' on instagram," The Atlantic, October 2018, accessed: 08.12.2019. [Online]. Available: https://www.theatlantic.com/technology/archive/2018/10/teens-face-relentless-bullying-instagram/572164/

[88] S. Hinduja and J. W. Patchin, "Offline consequences of online victimization: School violence and delinquency," *Journal of School Violence*, vol. 6, no. 3, pp. 89–112, 2007.

[89] S. Ho, "Girls report three times more online harassment than boys amid rise in cyberbullying," Time, July 2019, accessed: 08.12.2019. [Online]. Available: https://time.com/5636059/girls-online-harassment-cyberbullying-report/

[90] S. Hinduja and J. W. Patchin, "Connecting adolescent suicide to the severity of bullying and cyberbullying," *Journal of School Violence*, vol. 18, no. 3, pp. 333–346, 2019.

[91] C. D. Marcum and G. E. Higgins, "Examining the effectiveness of academic scholarship on the fight against cyberbullying and cyberstalking," *American Journal of Criminal Justice*, pp. 1–11, 2019.

[92] J. Snakenborg, R. Van Acker, and R. A. Gable, "Cyberbullying: Prevention and intervention to protect our children and youth," *Preventing School Failure: Alternative Education for Children and Youth*, vol. 55, no. 2, pp. 88–95, 2011.

[93] S. Hinduja and J. W. Patchin, "Preventing cyberbullying: Top ten tips for educators," Cyberbullying Research Center, June 2018, accessed: 08.12.2019. [Online]. Available: https://cyberbullying.org/Top-Ten-Tips-Educators-Cyberbullying-Prevention.pdf

[94] ——, "Cultivating youth resilience to prevent bullying and cyberbullying victimization," *Child Abuse & Neglect*, vol. 73, pp. 51–62, 2017.

[95] E. Siner, "Facebook takes on cyberbullies as more teens leave site," NPR.org, November 2013, accessed: 08.12.2019. [Online]. Available: https://www.npr.org/sections/alltechconsidered/2013/11/07/243710885/facebook-takes-on-cyberbullies-as-more-teens-leave-facebook

[96] J. M. van der Zwaan, V. Dignum, C. M. Jonker, and S. van der Hof, "On technology against cyberbullying," in *Responsible Innovation 1*. Springer, 2014, pp. 369–392.

[97] J. Juvonen and E. F. Gross, "Extending the school grounds? Bullying experiences in cyberspace," *Journal of School Health*, vol. 78, no. 9, pp. 496–505, 2008.

[98] Massachusetts Aggression Reduction Center, "Preventing cyberbullying: Top ten tips for educators," Bridgewater State University, https://www.marccenter.org/educator, 2018, accessed: 08.14.2019.

[99] D. Zapf and S. Einarsen, "Individual antecedents of bullying: Victims and perpetrators," *Bullying and emotional abuse in the workplace. International perspectives in research and practice*, vol. 165, p. 183, 2003.

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

[101] B. Love, B. Crook, C. M. Thompson, S. Zaitchik, J. Knapp, L. LeFebvre, B. Jones, E. Donovan-Kicken, E. Eargle, and R. Rechis, "Exploring psychosocial support online: a content analysis of messages in an adolescent and young adult cancer community," *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 10, pp. 555–559, 2012.

[102] P. Wisniewski, "The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience?" *IEEE Security & Privacy*, vol. 16, no. 2, pp. 86–90, 2018.

[103] O. Solon, "Teens are abandoning facebook in dramatic numbers, study finds," June 2018, accessed: 08.15.2019. [Online]. Available: TheGuardian,https://www.theguardian.com/technology/2018/jun/01/facebook-teens-leaving-instagram-snapchat-study-user-numbers

[104] Snap Inc., "Privacy centre - our privacy principles," Snap Inc., 2019, accessed: 08.15.2019. [Online]. Available: https://www.snap.com/en-GB/privacy/privacy-center/

[105] A. Lenhart, M. Duggan, A. Perrin, R. Stepler, H. Rainie, K. Parker *et al.*, *Teens, social media & technology overview 2015*. Pew Research Center [Internet & American Life Project], 2015.

[106] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[107] Y. Chen, Y. Zhou, S. Zhu, and H. 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*. IEEE, 2012, pp. 71–80.

[108] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, vol. 2. IEEE, 2011, pp. 241–244.

[109] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Computers in Human Behavior*, vol. 93, pp. 333–345, 2019.

[110] Tech vs Abuse (Comic Relief), "Tech vs abuse: Research findings," Tech vs Abuse, 2019, accessed: 07.28.2019. [Online]. Available: https://www.techvsabuse.info/research-findings

[111] P. Ekblom, "Designing products against crime," *Encyclopedia of Criminology and Criminal Justice*, pp. 948–957, 2014.

[112] R. V. G. Clarke, *Situational crime prevention*. Criminal Justice Press Monsey, NY, 1997.

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

[114] SaveLives, "For Maracs," accessed: 08.18.2019. [Online]. Available: http://www.safelives.org.uk/training/maracs

[115] M. Gerber and R. Von Solms, "Management of risk in the information age," *Computers & Security*, vol. 24, no. 1, pp. 16–30, 2005.

[116] S. Faily and I. Fléchais, "Barry is not the weakest link: Eliciting secure system requirements with personas," in *Proceedings of the 24th BCS Interaction Specialist Group Conference*. British Computer Society, 2010, pp. 124–132.

[117] ——, "The secret lives of assumptions: Developing and refining assumption personas for secure system design," in *International Conference on Human-Centred Software Engineering*. Springer, 2010, pp. 111–118.

[118] G. Sindre and A. L. Opdahl, "Eliciting security requirements with misuse cases," *Requirements engineering*, vol. 10, no. 1, pp. 34–44, 2005.

[119] R. Stevens, D. Votipka, E. M. Redmiles, C. Ahern, P. Sweeney, and M. L. Mazurek, "The battle for new york: a case study of applied digital threat modeling at the enterprise level," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 621–637.

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

[121] R. W. Reeder, I. Ion, and S. Consolvo, "152 simple steps to stay safe online: security advice for non-tech-savvy users," *IEEE Security & Privacy*, vol. 15, no. 5, pp. 55–64, 2017.

[122] S. Parkin, A. Van Moorsel, P. Inglesant, and M. A. Sasse, "A stealth approach to usable security: helping it security managers to identify workable security solutions," in *Proceedings of the 2010 New Security Paradigms Workshop*. ACM, 2010, pp. 33–50.

[123] L. Reinfelder, R. Landwirth, and Z. Benenson, "Security managers are not the enemy either," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 433.

[124] M.-E. Paté-Cornell, M. Kuypers, M. Smith, and P. Keller, "Cyber risk management for critical infrastructure: a risk analysis model and three case studies," *Risk Analysis*, vol. 38, no. 2, pp. 226–241, 2018.

[125] A. A. Ganin, P. Quach, M. Panwar, Z. A. Collier, J. M. Keisler, D. Marchese, and I. Linkov, "Multicriteria decision framework for cybersecurity risk assessment and management," *Risk Analysis*, 2017.

[126] I. Fléchais, C. Mascolo, and M. A. Sasse, "Integrating security and usability into the requirements and design process," *International Journal of Electronic Security and Digital Forensics*, vol. 1, no. 1, pp. 12–26, 2007.

[127] J. A. Dykstra and S. R. Orr, "Acting in the unknown: the Cynefin framework for managing cybersecurity risk in dynamic decision making," in *2016 International Conference on Cyber Conflict (CyCon US)*. IEEE, 2016, pp. 1–6.

[128] T. Islam, I. Becker, R. Posner, P. Ekblom, M. McGuire, H. Borrion, and S. Li, "A socio-technical and co-evolutionary framework for reducing human-related risks in cyber security and cybercrime ecosystems," in *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Springer, 2019, pp. 277–293.