A cyber-risk framework for coordination of the prevention and preservation of behaviours

Parkin, Simon; Chua, Yi Ting

DOI
10.3233/JCS-210047

Publication date
2022

Document Version
Final published version

Published in
Journal of Computer Security

Journal of Computer Security 30 (2022) 327–356
DOI 10.3233/JCS-210047
IOS Press

A cyber-risk framework for coordination of the prevention and preservation of behaviours[1]

Simon Parkin [a,*] and Yi Ting Chua [b]
[a] *Delft University of Technology, Delft, Netherlands*
E-mail: s.e.parkin@tudelft.nl
[b] *University of Alabama, Tuscaloosa, AL, United States*
E-mail: ychua@ua.edu

Abstract. Cybersecurity controls are deployed to manage risks posed by malicious behaviours or systems. What is not often considered or articulated is how cybersecurity controls may impact legitimate users (often those whose use of a managed system needs to be protected, and preserved). This oversight characterises the 'blunt' nature of many cybersecurity controls. Here we present a framework produced from consideration of concerns across methods from cybercrime opportunity reduction and behaviour change, and existing risk management guidelines. We illustrate the framework and its principles with a range of examples and potential applications, including management of suspicious emails in organizations, and social media controls. The framework describes a capacity to improve the precision of cybersecurity controls by examining shared determinants of negative and positive behaviours in a system. This identifies opportunities for risk owners to better protect legitimate users while simultaneously acting to prevent malicious activity in a managed system. We describe capabilities for a novel approach to managing sociotechnical cyber risk which can be integrated alongside elements of typical risk management processes. This includes consideration of user activities as a system asset to protect, and a consideration of how to engage with other stakeholders in the identification of behaviours to preserve in a system.

Keywords: Risk management, cyber risk, sociotechnical security

1. Introduction

Cyber-risk controls are deployed within a managed IT system, such as in a business or an online service platform, to manage cyber risks and address unknown or anticipated malicious behaviour. Implicit in common security and privacy risk management practices is that if a control is well-intentioned, it will not do any harm to those people and activities it is meant to protect. Cyber threats can impose a range of different harms upon legitimate users [5], however so can cybersecurity risk controls if not carefully considered [25].

These harms can range in their impact. This can include additional effort imposed upon end-users for everyday security procedures, such as login processes for registered users, or personal online banking.

[1] This paper is an extended and revised version of a paper presented at the International Workshop on Socio-Technical Aspects in Security.

[*] Corresponding author. E-mail: s.e.parkin@tudelft.nl.

Harms can also include legitimate users being removed from a system, or their activity being wrongly classified as malicious.

The critical trait of an unintended harm in a managed system is that legitimate users may be expected to use the system at the same time that its design prevents them from accessing it, as their behaviour shares aspects with the behaviour of a known or perceived threat to the system. Such *unintended harms* may be more severe for specific user groups who lack targeted support (such as the technical skills assumed to follow basic advice), or are inadvertently treated as malicious entities (e.g., by rules for identifying suspicious activity on a social media platform). An example would be previous admissions that facial recognition technology in active use 'lacks the diversity it needs' to avoid biases in automated decisions [11].

The potential for risk controls to harm legitimate users is pronounced in modern IT systems. The *hyperconnectivity* these systems embody [65] means that malicious and legitimate human activity in the same IT environment can have some of the same observable behaviours and use of the same infrastructure. For example, they can access online accounts through the same interface.

There is a need to ensure in advance that a candidate risk control does not impact the existing activities of legitimate users. In the examination of unintended harms of cyber-risk controls, what must also be considered is that IT supports a vast array of day-to-day activities. To effectively protect IT systems there is an increasing need to understand and safeguard how IT is used for productive activities, rather than focusing only on securing the IT systems.

Various methods exist for analysing a whole system to discourage a malicious behaviour (e.g., [27, 29]), or to promote and encourage positive behaviours [77] (Section 2). We consider the latter schools of science together, as a means to avoid 'blunt' controls which reduce malicious behaviours and also impact legitimate behaviours, particularly if selective user groups or behaviours are inadvertently treated as if they are malicious. An example would be changing system features to stop an attack, but making other benign activities difficult or impossible with the same action. This is the case with Facebook blocking a photographer's advert image of fireworks because of a judgement that it was promoting weapons [12].

To our knowledge, the interplay between these two groups of approaches has not been considered within cyber-risk management, though formative and disparate activities can be found (Section 2.4). This consolidation leads to new approaches to (i) conceptualising user activities as sociotechnical assets (Section 3), and (ii) addressing the sociotechnical *precision* of cyber-risk controls, to target only malicious or unwanted behaviours (Section 4).

1.1. Motivation and contribution

The work here progresses a number of complementary aims. By addressing these aims together, we produce guidance for the protection of legitimate use of a managed system/service alongside *concurrent* prevention of malicious use.

- Identify complementary capabilities from the domains of crime prevention and behaviour change support, toward identifying directions for the management of both malicious and legitimate user behaviours in the same environment. The aim here would be to reduce unintentionally creating a system or service that acts to prevent legitimate use for particular legitimate users and use conditions in the process of providing security;
- Characterise *interference* between malicious and legitimate activity in managed environments which can potentially lead to unintended harms. We focus on alignment of a system-level view of a managed security infrastructure, as would be accessible to cyber-risk owners, to the perspectives of

stakeholders who use the environment. This involves exploring shared determinant factors between malicious and legitimate behaviours, elements which our proposed framework brings together;
- Add to accepted cyber-risk management approaches to encourage cyber-risk owners to reduce the possibility for unintended harms stemming from the controls they deploy. This would promote advanced risk management capabilities using familiar tools, which is especially important given the need to consider complex sociotechnical interactions. Within this, we address gaps in existing risk management approaches to explicitly consider legitimate user behaviour as an asset to protect.

Together, these elements are a foundation for holistic cyber-risk management which is "user-friendly while abuser-unfriendly" [39]. We apply the proposed framework to case studies on phishing in organizations and abusive behaviours on Social Media Platforms (SMPs), where there are many cross-cutting concerns in each example (Section 5). We examine limitations to the approach, future directions, and issues in cyber-risk management that the framework exposes in the discussion (Section 6) and close with conclusion and next steps (Section 7).

2. Managing security for an ecosystem of behaviours

With IT systems underpinning so much of what people do in their normal lives, legitimate users and malicious actors are using the same infrastructure and technologies, making it more difficult to distinguish between their activities. To address this, we differentiate between different kinds of behaviour within an environment (Section 2.1). This is framed from the perspective of whether behaviours in a managed IT environment should be regarded as legitimate or malicious, and in turn how they would ideally be treated based on their classification. We then explore the complementary research areas of crime science and crime prevention (Section 2.2), and behaviour change science (Section 2.3). It is in this exploration that we highlight comparable features of these two areas, considered alongside current capabilities within information security management (Section 2.4). We provide examples which contrast 'blunt' and 'precise' approaches to security management (Section 2.5), and a review of related work (Section 2.6).

2.1. How secured systems are used

User-facing controls may restrict or direct behaviour, such as corporate log-in systems, or browser warnings which encourage users away from web-pages. Controls can also include policies or advice which users are expected to follow. Here we look at how legitimate users may be inadvertently treated as if malicious. This differentiates from malicious or wilfully careless users considered legitimate by the system, and latent conditions (here, 'harmful' security design) and unwitting mistakes (or 'active failures') by legitimate users when enacting behaviours [88].

We refer to a legitimate user not in terms of any legal premise, but as determined by the configuration of a system or service. That is, if the terms of a service allow a user to access the service, security controls should support – rather than impact (hinder or obstruct) – that use. Here, legitimate behaviour is activity by users that a system or service is intended to serve as defined in its configuration. Such behaviours are the legitimate behaviours presumed 'positive' to have in a system, that the system or service aims to promote; otherwise, they would be regarded as malicious, with efforts directed to prevent those behaviours.

We refer to *cyber-risk owners* as the stakeholders in an IT environment who have the authority and decision-making responsibility to enact changes to the cybersecurity apparatus within that environment. This apparatus includes technical and sociotechnical controls.

We also directly address an aspect shared with unusable systems, that the controls appear to be difficult for users to interact with. Here our focus is primarily where difficulties or obstacles occur when an intended, legitimate user shares traits with what the system regards as a malicious actor – the controls intend to counteract the malicious actor but also prevent the legitimate user from making use of the system. When this occurs, legitimate use is slowed or prevented as in a less-than-usable system, yet the legitimate user is expected to still use the system. Examples range from needing to provide complex credentials to access a system (which a user who has access may find too complex to recall or provide), to not being able to access a service because it routinely classifies the user's activity as being suspicious (e.g., blocking uncharacteristic banking transfers or payments).

Two difficulties emerge in establishing the prevalence of unintended harms of cybersecurity [25]. Firstly, they may affect only a small part of the population of legitimate users. Secondly, harms may manifest in a way that is represented in visible systems in subtle ways (from the perspective of the system owner) even if great harm is experienced by the user. We presume that most controls work for a 'general case' that the control designers and system managers have in mind. Harms may be invisible to system owners, due to the aforementioned comparatively small population of impacted users, but also because users may be displaced, their behaviour prevented or distorted, etc. [25]. Fundamentally, there is little existing capability to measure unintended harms imposed upon legitimate system users.

Below we differentiate between categories of legitimate and malicious behaviours. Considering the two difficulties mentioned above, these behaviours may manifest to varying degrees in an active system.

- **Legitimate and anticipated behaviour.** In most cases the legitimate behaviour that the system is envisioned and designed to support is able to proceed. If this were not possible, it would become very obvious that the system was not working. A system not working for many would likely be noticed in the same way as a power or service outage [13], where users complain or remark on e.g., social media, and organizations and news entities noticed it almost immediately. Where this relates to security is if many service users experience issues due to a wide-reaching mismatch of security and privacy features with the expectations for typical use. A particular case is where users realise the anticipated behaviour and have options to move to an alternative service. An example of the latter is the announcement of changes to WhatsApp service monitoring (which was seen for most users as being the publicising of service terms which already applied [14]).
- **Legitimate but difficult behaviour.** This relates most closely to existing research directions around usable security [94], where security makes the goals of the primary, productive task more difficult to reach. Unusable controls can include additional security measures to an existing process in a way that overlooks added burden on all users, such as being required to provide additional credentials to access an online/networked service. This can prompt *workarounds* akin to behaviours in the next category, which may look like avoidance or erosion of security controls.
  An imprecise response to such erosion of control may be to more rigidly control the system – this makes the already difficult behaviour even more difficult, and potentially impossible. Imprecision can also lead to the *lack* of user activity, i.e., users deciding the service is too difficult to use and migrating to a comparable alternative, if they are able or empowered to do so (where further harms arise if they are not) [25].
- **Legitimate but not anticipated behaviour.** Risk controls may be flexible enough that users can themselves take action to make them easier to use. If these workarounds are used persistently over

time, these become *coping strategies*. Flexibility also allows *appropriation* of technologies, where users would make a technology better fit their own practices [84]. One example of a behaviour that may not be anticipated but still possible would be the sharing of credentials or sensitive information between legitimate users, to carry out a legitimate behaviour within this system, such as giving personal bank card details to a family member so that they can make purchases on their behalf [73]. Coping strategies such as reusing a password across multiple services also fall under this category. This is where the service was designed with a particular view of 'user' in mind, which is not nuanced enough to capture everyday uses. This category is then most likened to being a 'gray area'. A harm here would be to force legitimate users to conform to the service's strict view of the user, rather than the service accommodating the user. Another service, or a future feature of the same service, may do a better job of accommodating such behaviours, but only if existing barriers and harms have been understood and removed.

- **Legitimate behaviours with unanticipated needs.** This refers to user groups within the system that may be vulnerable or have specific expectations. Some examples include the needs of users with disabilities or impairments [89,104], or older or less experienced users [70]. A comparable example outside of cyber risk would be increased provision of accessibility options in modern video games [15], which security controls could seek to emulate, so that there are alternative technologies to meet varying needs of users.
- **Malicious but feigning legitimate behaviour.** This is a problem case outside of our goals, where this class of behaviours refers to instances of malicious actors mimicking legitimate behaviour within a system; the malicious intention may be exposed or flagged at some later point in time.
  Rules may be put in place in the system as an effort to identify these types of malicious behaviour, based on an approximation of the behaviour. One example is online romance fraud [37], wherein malicious actors, using fake identities, purport to be legitimate users on dating platforms, to initiate relationships as a means to obtain financial gains [106]. Actions to manage these kinds of behaviours become an issue if risk owners act on what they believe to be a distinguishing feature of only the malicious actor, which is also shared with legitimate users, e.g., monitoring or requiring additional credentials from users purporting to be from a particular location or platform.
  Harms may be experienced by legitimate users if they must absorb any impact or damages which arise from controls aimed at malicious behaviour. A stark example that demonstrates this is when one US city government was prevented from rapidly registering temporary email addresses in an effort to recover from a ransomware attack, wherein the services they were using categorised this as activity normally associated with a spam campaign [81].
- **Malicious behaviour here but not somewhere else.** This is where there are variances in the cyber-risk controls and policies across otherwise comparable services with the same primary functionality. Restrictions to behaviour at one location (i.e., social media platform) can lead to the migration of restricted activity (and users) to another platform. Such behaviour coincides with the concept of crime displacement [95]. One recent example of this type of behaviour is the migration of users from mainstream media to alternative platforms such as Gab and Parler, due to moderation of mainstream platforms by companies such as Amazon [92].

A note on the above is that security is often *added to* an existing system or process that serves a productive task. This points to regular (cyber) risk management activities as opportune moments to review how well the cyber-risk controls for a system fit with legitimate user behaviours while also preventing malicious behaviour (see Section 2.4). There must be a mechanism for engaging with the needs of legitimate users to understand whether a control has introduced or is going to introduce irrecoverable

harms. This highlights the need to view legitimate behaviours as 'assets' of the system, which ought to be encouraged and protected alongside security-related measures.

2.2. Discouraging malicious security-related behaviours

Scholars have explored the applicability of existing theoretical frameworks and approaches from crime prevention to the domain of cybercrime. Both social learning theory and general theory of crime have been applied to examine cybercrime, such as hacking behaviours [18,74,75], where both theories focus at the level of the individual.

Other crime prevention approaches focus on the opportunity structures and immediate environment as causes of criminal acts. *Situational crime prevention (SCP)* has shown success in addressing online crimes such as data breaches [30]. SCP is a framework of strategies with aim to reduce criminal opportunities arising from the immediate environment [26,27]. Rather than viewing crime as a result of criminal predispositions, it views crime as the result of one's deliberate choices and decisions [26], affected by a person's immediate situation and circumstances. This shapes the three inter-related features of *SCP*, being specificity of the crime, the immediate environment, and the individual's perception and decision to commit a malicious act [26,27]. Together, these three features enable researchers to identify viable points of intervention and prevention for criminal acts. Malicious and legitimate activities share use of the same networks and online services [64], such that a malicious user's circumstances, behaviour decisions, and environment must be considered alongside those of legitimate users.

These intervention and prevention techniques reduce criminal opportunities through identifying potential components that can be changed. There are currently 25 techniques falling under five categories, each containing five techniques: 1) increasing efforts, 2) increasing risks, 3) reducing reward, 4) removing excuses, and 5) reducing provocation [27,32,108].

Routine Activity Theory (RAT) emphasises the circumstances around when crimes occur [29,51]. Its main proposition is that crime occurs as the convergence in space and time of a suitable target, a likely offender, and the absence of a capable guardian [29,51]. In addition, the convergence of these elements is further dictated by the spatial and temporal patterns of community structure consisting human interactions and relationships [29]. Thus, normal, daily and/or recurring activities within a community determine the distribution of criminal opportunities because these activities affect when, how, and where the three elements of crime would converge. The development of RAT then informs that prevention of crime interacts with – and must consider – the activities of legitimate individuals.

The capable guardianship element refers to any person with the potential capabilities to prevent the occurrence of a crime [29]. The concept of guardianship was later expanded to identifying agents capable of discouraging crime within each component [50]. Despite such expansion, Hollis and colleagues [60] highlight the variations in the conceptualisation and measurement of the concept in scholarly work, such as confusion between guardianship and target hardening (which addresses suitability of targets) as well as guardianship and social control (both formal and informal). As a result, Hollis and colleagues [60] propose the following definition: *Guardianship can be defined as the presence of a human element which acts – whether intentionally or not – to deter the would-be offender from committing a crime against an available target.* What we emphasise in our framework is the need to deter cyber-related offences – that guardianship is worthwhile – but that it must not impact the capacity for intended users to enter and use the managed system. Fundamentally, rather than protecting a 'system', we instead assert that there is user behaviour which must be guarded against harm and limitations to access.

RAT has been adapted to explain victimisation as a result of online lifestyle and routine behaviours, while conceptualising computer and cybersecurity features as effective guardians [23]. However, when

applying *RAT* and *SCP* to cyberspace, it raises the issue of contact between offenders and targets or victims, due to the nature of online interactions. Reyns [91] utilizes the concept of *system problems* while examining cyberstalking using the *SCP* framework. System problems denote that the offenders and victims can be connected via networks, and the convergence through networks suffices as a condition for crime to occur. In addition, these networks function much like places, and crime prevention should involve place managers such as network managers [36]. We focus on risk owners within a managed IT infrastructure as 'guardians' of legitimate users in a system, acting to reduce the opportunities and capacity to conduct malicious activity.

2.3. Preserving positive security-related behaviours

Referring to the different forms of behaviour in a managed system (Section 2.1), behaviour change approaches not only inform how to change a behaviour, but also the conditions which must be in place to *maintain* a behaviour [28]. This is important as a risk owner in a managed networked system must not take action that impairs an existing legitimate behaviour that is ongoing in the system. Our focus is on how behaviour change approaches inform ways to preserve existing behaviours. Unchecked action that focuses solely on reducing malicious activity in a networked environment has the potential to undo the conditions of an existing legitimate behaviour; in terms of behaviour change, this makes the behaviour more difficult, thereby de-emphasising it and making it less viable as a behaviour to enact [28].

Under this premise, cyber-risk management as relates to behaviours is often regarded as furnishing legitimate users with the controls and protections they need to use a service or system effectively. There is a risk that usability challenges are imposed upon users if a cyber-risk solution is inappropriate or awkward to use [94]. Usability challenges may be imposed if initiatives 'harden targets' (as in Section 2.2), without considering the need to act as a guardian and preserve legitimate behaviours. If the duty of guardianship is overlooked, one example of a harm introduced to the system is when browser warnings misclassify legitimate websites as dangerous and dictate that users avoid visiting them [48].

A range of factors are critical to encouraging an individual to adopt a positive behaviour. The COM-B model [77] distills critical factors for promoting behaviour change, namely capability, opportunity, and motivation. Similarly, the 'B = MAP' behaviour change framework [53] encompasses the need for a combination of Motivation, Ability, and Prompt for new behaviours to form. Prompts, as similar to opportunities, have been explored for security elsewhere (e.g., security advice for consumers [86]). Both frameworks consider the role of the environment and other stakeholders in enabling behaviours; an individual cannot enact a behaviour if the environment seems to be working against them.

Intervention Mapping [9], within the health domain, highlights a need to support specific outcomes for a behaviour, and to precisely target the underlying determinants that enable the behaviour to happen, especially for specific subgroups. Being precise is then framed as key to encouraging and sustaining good behaviours. These principles have been applied in targeting cybersecurity awareness initiatives [90]. Also within the health domain, the PRECEDE-PROCEED intervention framework [54] emphasises the development of targeted interventions for a particular group and behaviour, including attention to factors which promote or prevent a behaviour. Here we argue that the need for such precision should be similarly emphasised in the design of cybersecurity interventions.

We focus on where the risk owner is in 'agreement' with the service owner about what the service is, but has an underdeveloped understanding of how users experience use of the system. It is then imperative to understand how users are *enabled* to use a system, and in turn where cyber-risk controls hamper their access while also proposing to be there to help them. An example would be automated detection algorithms in the context of harmful content on social media, which may flag legitimate content.

2.4. Risk management for systems of behaviours

The literature in both crime reduction and behaviour change (for our purposes, behaviour preservation) share a common ground: factors in the environment, individual characteristics and motivations, and patterns of behaviour. These factors need to be viewed from the context of current approaches to managing security-related behaviours. We therefore refer to cyber-risk management literature aimed primarily at organizations and large networked systems. This allows us to build on practices familiar to risk managers.

Various risk management approaches have hinted at issues tangential to our aims, albeit without directly addressing the linked impacts between efforts to *prevent* and *preserve* different IT-facilitated behaviours concurrently. ISO/IEC 27005:2011 ('Information security risk management') [66] explicitly includes 'Identification of consequences', though focusing on the consequences of a threat upon an asset, with no explicit examination of the impacts a control may have upon that asset. The broader ISO/IEC 31000:2018 'risk management' guidelines [20] acknowledge that risk management efforts may produce unintended consequences, noting that implementation of risk treatment plans ought to ensure that controls are effective when they are deployed, or otherwise that any risks they introduce are managed.

Related 'Risk management techniques' in ISO/IEC 31010:2009 [62] outline *consequence analysis*, to capture impacts including those affecting different objectives and stakeholders. It is also advised to capture how consequences relate to the original objectives, and secondary consequences, with further consideration of *hazards*, including physical harm. The potential for knock-on impacts from managing one risk upon another risk are highlighted, but not further developed. The need to ensure a 'freedom from risk' is acknowledged in the digital domain within standards for software development (as in ISO 25010 [63]). Techniques exist in cyber-risk management standards which can minimise unintended harms to legitimate users, but are not being coordinated to do so.

The NIST 'Risk Management Framework for Information Systems and Organizations' standard [67] brings attention to "potential adverse effects on individuals", and that some capabilities must be upheld to meet stakeholder needs. Our framework addresses a need for *existing* security and non-security capabilities to escape impact from subsequent countermeasures. The OCTAVE risk management process [6] considers how a risk management strategy itself can impact 'exposed assets'. We argue that users and behaviours linked to known, permitted capabilities within a system should be explicitly regarded as assets to protect, echoing directions outlined by a successor to OCTAVE, OCTAVE Allegro [21].

Standards such as those discussed above are not followed widely by all cyber-risk managers [78]. We regard these standards as being representative of practice, as cyber-risk owners may apply a mix of standards and self-developed 'folk risk analysis' techniques.

We do not regard cyber-risk controls as shaping user behaviours to fit the control, but instead address the need to shape cyber-risk controls to distinguish between legitimate user behaviours and malicious behaviours. Improper use of infrastructure by legitimate users is a usability and awareness issue. Where crime prevention meets society, there may be efforts to differentiate 'what we want less of' and 'what we want more of' [107]. We differentiate between 'what we want less of' (malicious behaviour) and 'what we want to retain opportunities for the same or more of' (legitimate behaviours). This requires knowledge of the malicious and legitimate behaviours already in the system and how they interact with cyber-risk controls. The three pillars of the COM-B model – capability, opportunity, and motivation – serve as a simplified bridge between efforts to reduce malicious behaviours and efforts to preserve legitimate behaviours.

2.5. Existing examples

The following are examples of where consideration of the interplay between malicious behaviours and legitimate user activities has resulted in precise targeting of negative outcomes while preserving positive behaviours.

- **Phishing reduction through token authentication.** Google employees were provided with two-factor authentication (2FA) tokens [72]. Rather than relying solely on training to avoid phishing attacks, this recognises that email links and service access can be typical in work, and that malicious/fake links etc. may be difficult to spot all of the time, making them difficult to separate. By using physical tokens to enable system access, a 'successful' phishing attack does not gain enough credentials to compromise a system (nullifying the value of knowledge-based credentials). This also means that employees are not under pressure to identify malicious links themselves to avoid compromise at all cost, and as a result warp their treatment of legitimate emails.
- **'Loan-phones' during digital forensics activities.** When a personal phone is being analysed for evidence of domestic abuse, some police forces in the UK provide a temporary phone, while some may not (which can factor in grave consequences [82]). A temporary phone preserves a person's capacity to reach their social support network or seek help. Here, a control to collect data of malicious activity (from smartphones) inadvertently removes the smartphone from its user; provision of loan phones reduces the impact to positive behaviours.
- **Socio-technical password controls.** There have been approaches in UK policy[1] to shift effort in managing passwords from end-users to background technical controls, so that legitimate users do not face the same difficulties that are created to dissuade malicious behaviour. For instance, system monitoring may be able to detect suspicious system activity and block access to legitimate login sites. 2FA tokens, as above, is a similar measure, reducing the heavy reliance on legitimate users to protect their passwords.

2.6. Related work

The SCENE framework [33] suggests to develop cybersecurity behaviour change options so that the most secure options are most accessible, ideally as 'defaults' (as applied for wi-fi selection [103]). Similar to behaviour change and crime reduction approaches, SCENE advocates co-creation of solutions with target audience and stakeholders. We posit that the available options for using IT securely may be reduced by efforts to reduce malicious activity.

Agrafiotis et al. describe a taxonomy of *cyber harms* [5] which may be observed in organizations. The taxonomy comprises five broad themes, including digital harm, and social and societal harm. The authors posit that analytical tools are necessary to reduce these harms, and as part of risk assessment. Similarly, Chua et al. [25] encourage risk managers to explore the potential for *unintended harms* to emerge as a result of their own risk controls. The authors' framework emphasises the need to support vulnerable populations who may experience harms if risk controls work against them rather than for them. We identify factors which contribute to unintended harms, rather than consequences.

The Security Function Framework (SFF) [38] surfaces design considerations for sustainable crime reduction solutions, and creation of new products. Ekblom notes that malicious actors and their (potential) victims may have *script clashes* [43], with a need to design solutions to "favour the good guys"; these

[1] "Password policy: updating your approach": https://www.ncsc.gov.uk/collection/passwords/updating-your-approach.

aims are further considered in the Vibrant SFF (VSFF) [107], which not only 'favours' good behaviours, but considers how to design societal interventions which both promote a good behaviour and reduce a bad behaviour at the same time. Where a crime reduction solution has a *niche* [40] in how it relates to "other products, people and places in the human, informational and material ecosystem", we pursue a similar notion of *precision*. As we consider user communities in IT ecosystems, this involves users, user behaviours, and infrastructure.

3. Foundational conditions

Risk management standards do not sufficiently articulate and address the needs to protect users and existing user behaviours. The consideration of legitimate user activities must be strengthened to match the level of capability *presented* in related risk management measures to prevent malicious behaviours. Managed systems must be attuned to the behaviour of their users more than is acknowledged in current approaches. There is almost no proactive effort at present to understand legitimate uses of a system *which are not a security concern*, as if they are separate from keeping a system secure. Cyber-risk controls must protect a *system of behaviours*: security measures should not impede these non-security behaviours, and practice can no longer rest on the assumption that security measures act in isolation from non-security activities. This motivates us to improve on current approach via an interdisciplinary, exploratory framework, which requires specific conditions to be met prior to application. These conditions include (a) surfacing stakeholder interaction in the roles and responsibilities of risk owners; (b) inclusion of legitimate/positive behaviours in cyber-risk management, and; (c) recording the dependencies between protected assets and behaviours.

3.1. Condition one: Stakeholder interaction in the responsibilities of cyber-risk owners must be surfaced

The identification and involvement of stakeholders in shaping controls is open-ended in current risk management approaches. Risk management standards are generally detailed in determining how the actors and constituent elements in a system may be adversely affected by an incident or malicious activity, but this same rigour is not applied to the effects of controls themselves. Where ISO 27005:2011 [66], for instance, refers to the 'scope and boundaries' for the malicious activity managed by a risk control, the notion of 'boundaries' in cyber-risk management requires development in terms of how user needs are identified with stakeholders and insulated from unintended effects stemming from cyber-risk controls. Techniques may be adapted relating to guardianship in *RAT*, or crime preventers and promoters in the work of Ekblom [40].

We make a simplifying assumption that a cyber-risk owner is afforded a more direct view of candidate risk controls and their features than any other stakeholder. Building on this, a cyber-risk owner is optimally positioned to look 'under the hood' at the causal factors of malicious behaviours that are managed by a control [98]; these causal factors may be shared with a legitimate behaviour. *Engagement with stakeholders* is encouraged to reach effective solutions, in both crime reduction and positive behaviour change. We then focus on those mechanisms under the view of a cyber-risk owner, which have the potential to impact other parts of the system.

We posit that the role of a cyber-risk owner should be augmented to have some responsibility for preserving the legitimate user behaviour they are aiming to protect, as a guardian. At present the pre-

vention of malicious behaviours is a primary aim, but this loses sight of the behaviours to be preserved. Current risk management standards mask these challenges; standards prescribe engagement with other stakeholders in the system, but only insofar as checking that security plans do not interfere with existing security controls already active in the system, and that the other stakeholders approve of the intentions of new controls. We propose that the two roles of preventer (of negative behaviours) and protector (of positive behaviours) be afforded the same visibility in risk management practices. Otherwise, a cyber-risk manager is acting solely as a preventer, as if security is the priority of the system.

3.2. Condition two: Legitimate behaviours in cyber-risk management ought to be catalogued

The activities of legitimate users, which create the value in a service or environment, are *sociotechnical assets*. There is a pronounced gap in existing cyber-risk management approaches, where these sociotechnical assets are not directly considered, despite being represented in systems in the likes of user profiles, behaviour data, and system management decisions/rules which act upon them. Risk management is at present centred around data and artefacts of value as static assets, but the behaviour of legitimate users which produces those assets is not directly acknowledged and protected. Changes in how aspects of existing risk management approaches are emphasised can realise more holistic, user-centred outcomes.

To illustrate, we draw on the VSFF [107]. This framework highlights advances in crime mitigation, to give equal weight to positive improvements as is given to crime prevention, and act to manage these goals *simultaneously*. Where this approach considers the design of a single project or artefact, we pose that just as risk standards stipulate that each security control must not interfere with *existing* security controls, it must also not interfere with *existing* legitimate user activities. This is mirrored in the 'human-as-solution' approach of Zimmermann and Renaud [110], where no one security management activity can be conducted as if in isolation from the rest of the system.

To borrow from physical crime prevention, the 'Grippa clip' [42] is an example of considering how to discriminate between a risk and existing legitimate practices. This is a clip which bags and coats can be hung from in public spaces, which allows legitimate users and property managers to continue using the space, while keeping the space clear, and making theft of bags more difficult.

3.3. Condition three: Dependencies between assets and legitimate behaviours should be recorded

Risk controls in an IT environment potentially restrict behaviour, users, and infrastructure [25], in turn affecting actual user behaviour. A risk owner making decisions about IT-security is unlikely to have a direct view of how legitimate users are actively using the system; instead they see the data assets produced by that legitimate behaviour. There is then a lack of explicit acknowledgement of the connections between what would normally be considered assets to protect, such as data and systems, and the legitimate user activities that produce and use those assets.

To address this shortcoming, we adopt a mechanistic approach similar to that described by Hatleback and Spring [57]. With this, a behaviour can be an *indexed entity*, as a file or data, but also exist as an activity in a system, producing a visible phenomenon. An example would be a 'delete' function which exists as rules, but can also be enacted as an activity which is run within the system. We consider this mechanistic approach as useful for looking backwards from a visible asset to consider what system activity produced it, and in turn, what legitimate user behaviours contributed to it (as in Fig. 1).

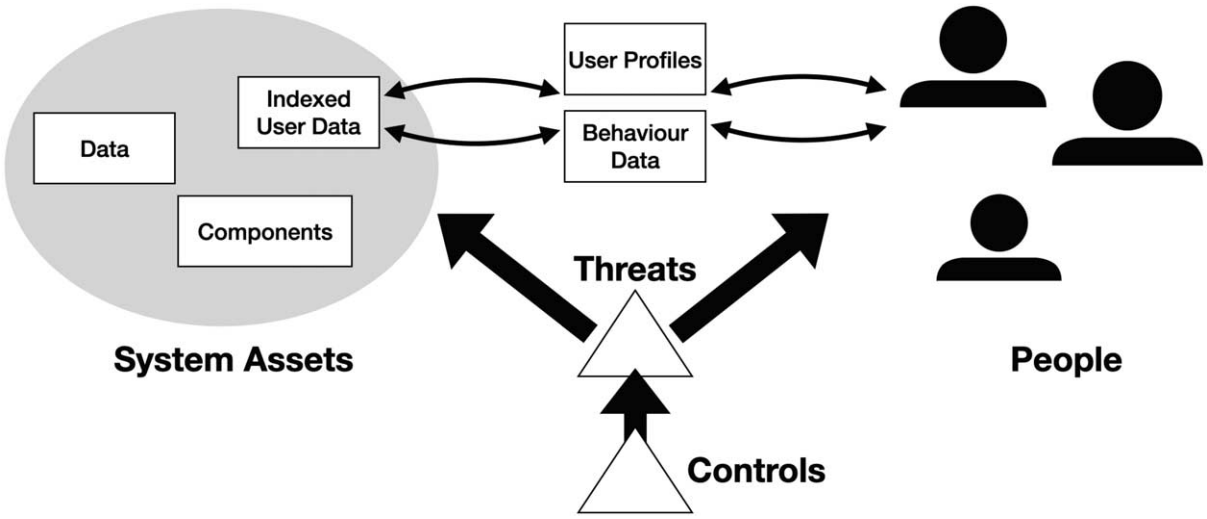

Fig. 1. Extending risk management artefacts to accommodate sociotechnical risk management. Individual people may interact with a system in such a way that user profiles and behaviour data are generated and maintained. These are then indexed user data, generated as system activities alongside the behaviours of people using a system.

A foundation for precision in sociotechnical security controls extends the definition of an asset to acknowledge the activities of users that contribute to the asset. This adapts the definition of indexed entities [57] to acknowledge the connection between the entity and the activity – that the activity *causes* changes to the entity (the data asset) – and that these changes are the phenomena that a cyber-risk owner can see. This relates (positive and negative) real-world behaviours to the identifiable data and IT systems which are subject to the decisions of a cyber-risk owner.

Critically, there is a feedback loop between System Assets and People (Fig. 1) – if there are rules about how data can be accessed and modified in a system, these rules may restrict the activities of People. Restrictions to what the entity can be and how the entity can be accessed can restrict what the activity producing it can be and who can use that entity. Examples include restrictions on credentials necessary to make a new account on a system, or checks for particular kinds of access to platform features which are permitted based on characteristics of a user's account.

At present there is no guarantee that the user activity which acts upon a protected asset is also being protected. Representing user behaviour leads to the challenge of coordinating two up-to-now distinct efforts. The first is removal of negative, malicious behaviours from the system (e.g., inflammatory posts on social media). The second is maintaining positive, legitimate behaviours already in the system (e.g., allowing users to share posts on social media). Where risk management often involves maintaining a *risk register* of top risks, a specific risk management activity is generally missing to address the second of these efforts; this would involve recording user behaviours which are active in the system and must be preserved, because they are contributing to assets. An example would be that a legitimate user from a particular geographic location should be able to make regular posts to a social media platform and share links if they would want to, but that malicious activity seeming to emerge from the same area, posting fake messages and sharing malicious links, ought to be stopped, as may happen in online romance scams [25]. The capacity to populate a *behaviour register* is needed, where this is a natural extension to existing risk management techniques, aligning with behaviour intervention approaches in Section 2.4.

4. Framework for precision in sociotechnical controls

In this section we describe our framework, which builds on approaches from crime prevention and behaviour change, towards an initial approach for identifying interference between features in a managed system which prevent or preserve user behaviours. We also discuss means to act on the precision of cyber-risk controls, and how a lack of precision may be measured.

4.1. Terminology

Before discussing the details of our proposed framework, we clarify the terms used to underpin it (as also in Fig. 2).

- **Positive Activities.** This refers to three of the four legitimate behaviours defined in Section 2.1 that we consider as 'good' (positive) behaviours to preserve within a system. Thus, positive behaviours can be (a) legitimate and anticipated, (b) legitimate but difficult (which should ideally be made less difficult but are not), and (c) legitimate but not anticipated (as behaviours which ideally would be accommodated if attention and resources are brought to them, with nothing otherwise that justifies excluding these users from the system). These behaviours may be associated with or exhibited by a Trusted User profile, or be observable in a Secure System; a behaviour that is legitimate but not anticipated may not manifest in a system unless action is taken to make it possible. Legitimate

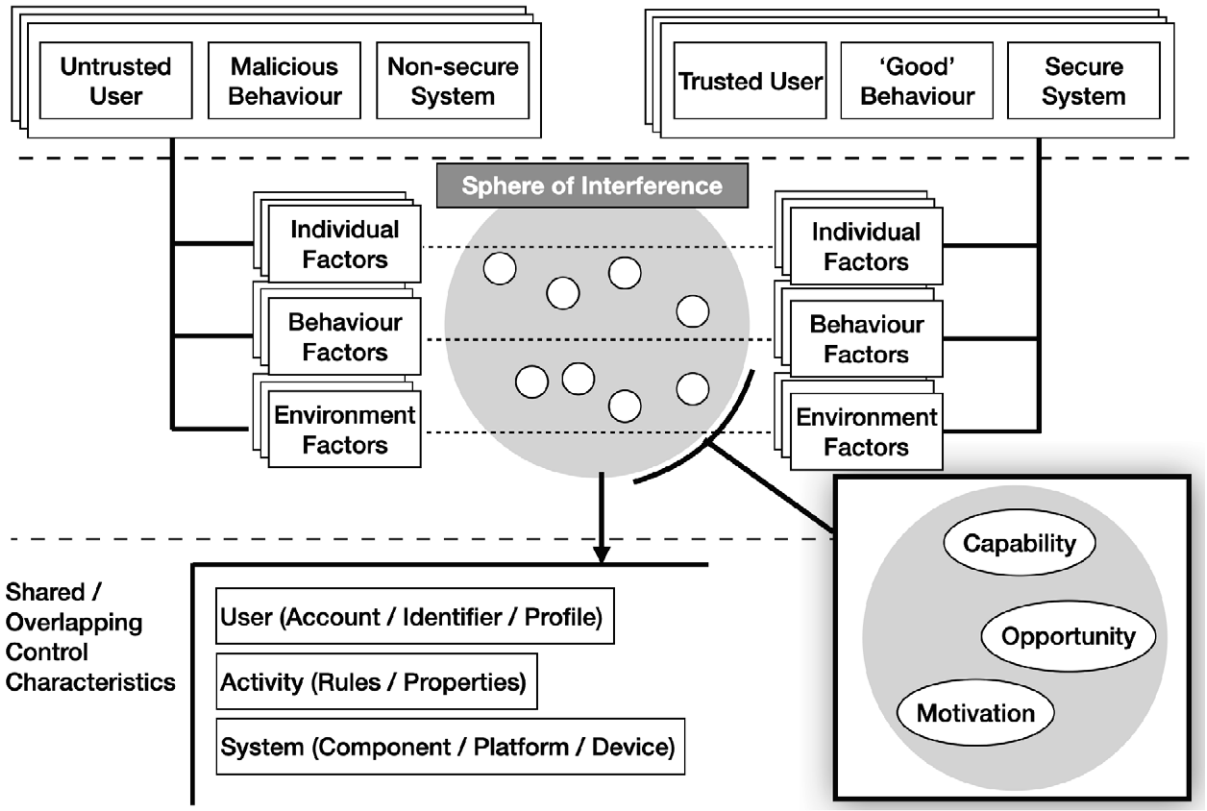

Fig. 2. Overview of interactions between negative and positive behaviours in a managed (cyber) system, and related controls.

behaviours and Trusted User profiles would ideally be precise enough so that they can be tracked in a 'behaviour register' (Section 3.3).

- **Negative Activities.** This refers to malicious behaviours as discussed in Section 2.1. Negative behaviours can (a) clearly fit the profile of a malicious behaviour, or become more difficult to discern, (b) be malicious but feigning legitimate behaviours, or (c) be malicious behaviour in one system but not another system. Malicious behaviour may be associated with broad definitions of Untrusted Users, or Non-secure systems.
- **Individual, Behaviour, and Environment Factors** are relevant to both types of behaviours within a system. These refer to factors which can be targeted or may be impacted by controls. These would be Individual Factors of the user, Behaviour Factors of what the user may do (as visible or detectable in the system), and Environment Factors of the system itself (such as interface or platform features). This is similar to Intervention Mapping, the PRECEDE-PROCEED framework, but also Situational Crime Prevention (SCP) (as discussed in Section 2). Any of these Factors may contribute to attributes shared between phenomena of a negative activity and phenomena of a positive activity. Where attributes are shared, this constitutes a Linkage.
- **Sphere of Interference.** This is essentially where a cyber-risk owner needs to unpick which phenomena, as seen in the data or analysis available to them from the managed environment, can be attributed to negative activity or positive activity (Section 4.1).
- **Control Characteristics.** If cyber-risk controls are not precise enough, there can be a range of properties in observed User Accounts (Untrusted User / Trusted User), Activity (Malicious Behaviour / 'Good' Behaviour), or Systems (Non-secure System / Secure System) which can be identified and refined, as in Section 4.4.

4.2. Intersection of behaviours to prevent or preserve

As in Fig. 2, we describe a method of sociotechnical cyber-risk management to coordinate refinement of precision in security controls. Existing (cyber)crime reduction techniques and behaviour change approaches amply describe how to manage individual behaviours. As a first step, we propose to consider the Capability, Opportunity, and Motivation of a Behaviour (COM-B) [77], as a shared terminology that represents aspects of both domains, to allow for comparison between two sets of specific behaviours and refinement of controls. A mediating set of terms and concepts serves as a *trading zone* [98] (or *translation zone*) between two domains of expertise, without requiring that one supersede the other.

Referring to the elements at the top of Fig. 2, as a premise, a cyber-risk owner will have an aim to prevent an Untrusted User from accessing a Secure System, or any user conducting a Malicious Behaviour. They may simply be aiming to reduce non-security in their system with no specific malicious user or activity in mind (by e.g., keeping up with the latest security patches available for their infrastructure). The cyber-risk owner will be aiming to protect a Trusted User, 'Good' (i.e., legitimate) Behaviours in the system, and prevent malicious use of or access to a Secure System.

Already here there is a potential disconnect which needs to be reconciled – the risk owner may act upon any one of the three malicious traits in the top-left of Fig. 2, and impact the complementary components in the top-right of the figure. An example would be blocking Untrusted Users, who are simply users not known in the system, from conducting Good Behaviours (e.g., a wave of new users joining a new social media platform from a country where it has just become available). This highlights that the security function may act on malicious activity in the environment, but does not directly govern the legitimate elements of the system (users, behaviours, or features). Yet it is the case that preventative controls can

impact these 'good' elements, as we see later with the consideration of *interference* between the two sides.

The intersection between system-views of malicious activity (upper-left of Fig. 2) and legitimate activity (upper-right of Fig. 2) illustrates where the connection between the two happens. As in Fig. 2, there can be a Sphere of Interference between identified Factors, where discerning between malicious and legitimate activity is not straightforward. As the two different views may be managed by different system/service stakeholders, it cannot be relied upon that they are using the same terminology.

The Sphere of Interference is subject to negotiations between both sides. We represent this by considering the behaviours that either side has in mind, represented objectively by the Capability, Opportunity (as designed or as occurs dynamically), and Motivation of behaviours. This is one example of a set of terms which can be used by both sides to articulate the behaviour they are acting to prevent or wish to preserve. A decoupling of Factors and COM-B elements is also useful here as the relationship between identified Factors and behaviours is not necessarily deterministic, especially with preventative measures. A cyber-risk owner may be acting proactively against particular Factors which they believe are linked to a particular behaviour (and its COM elements), e.g., that rapid registering of new email addresses with an email provider indicates only that they are going to be used for malicious purposes (where organizations impacted by ransomware have been seen to need to use the same tactic [81]).

By discussing the COM factors of behaviours [77] in the Sphere of Interference, it may be found that there are overlaps (depending on how specifically behaviours are defined, as discussed in Section 2). If differentiating between malicious and legitimate behaviours is difficult, this indicates where *linkages* between them are strongest, and the need to unpick them more critical, so as to avoid unintended harms to legitimate behaviours.

A further step is to identify sufficiently detailed definitions of User, User Behaviour, and Infrastructure. These are the elements arguably already familiar to a cyber-risk owner, but also which are closest to the configuration of an implemented technical control. These elements would also influence the COM factors in behaviours (as evidenced by risk controls preventing malicious behaviours). The extended asset definition in Fig. 1 supports this.

There is a separation of Factors and of Characteristics here primarily to make it clear that aspects of user activity may be discernible from system artefacts, but they are not a reasonable representation of user behaviours on their own. This is critical when considering 'soft' controls such as advice or policies. These would be represented as identified Factors to discourage or dismantle malicious behaviour or encourage to promote new secure behaviours. The framework also then serves as a means to explore if newly-proposed behaviours also interfere with existing behaviours; this is arguably a challenge in, for example, social media platforms, which encourage opportunistic social interactions while also discourage over-disclosure and interactions with untrusted contacts.

Crime reduction techniques (Section 2.2) are advocated here to identify negative behaviours. The behaviour change approaches in Section 2.3 are leveraged to identify positive behaviours to preserve. The latter requires a retrospective view of which behaviours are to be retained in the system, which is not exactly how behaviour change approaches are typically used, but indicates a need to catalogue behaviours much like there can be a record of the technologies deployed in an IT environment. Both approaches would interact with risk management approaches (Section 2.4) to identify candidate controls. In the case of positive behaviours, risk management approaches ought to be developed to effectively catalogue these behaviours as assets to protect.

4.3. Identifying lack of precision in risk controls

We regard a risk control as lacking precision if it imposes harm upon legitimate users in the service ecosystem, rather than a control which is not specific enough in targeting a particular risk; broad controls can come at a cost to users which share similar traits with attackers.

The framework illustrated in Fig. 2 prompts cyber-risk managers to explicitly decide how they are going to manage the potential for unintended harms upon legitimate users in their managed systems. The cyber-risk manager's responsibilities include target-hardening and acting as a guardian of legitimate users; they must decide to what extent they will proactively ensure that their control choices protect valuable assets and preserve legitimate behaviours.

From prior analysis in Section 2, our method includes the following steps. We include use cases of applying the steps of the framework in Section 5 to highlight the application of the framework in two scenarios: phishing in an organizational context and preventing online abuse on social media platforms.

Step 1. Record behaviours in the system.

1A **Identify active behaviour reduction activities.** This requires a catalogue of (malicious) behaviours being actively targeted, $R_1 - R_N$ (See both cybercrime reduction approaches in Section 2.2, and risk management approaches, Section 2.2). As in Fig. 2 and based on crime-reduction techniques, there is a need to consider here whether behaviour reduction activities or controls are acting to increase the effort of Capability, remove an Opportunity, or introduce some other deterrence effect such as increased costs, monitoring, etc. (affecting Motivation), to act to reduce or remove a behaviour from the service ecosystem. One example would be to introduce two-factor authentication to undermine the purpose of password-guessing attacks on login systems.

1B **Identify active behaviours to be preserved**. This set, $P_1 - P_N$, includes behaviours being promoted as part of active intervention programmes. This requires communication with other stakeholders in the system, as in common behaviour change approaches (Section 2.3). In organizations, the extraction of permitted behaviours can begin with access control policies, computer fair use policies, and include discussions with team managers to understand regular work activities [68]. In IT environments more broadly, this requires discussions with user representatives and local community experts (as with responding to tech-abuse [85]). Here we do not consider behaviour by legitimate users which will in the near future be identified as against a policy. This framework is not another means to detect malicious behaviour, but instead a proposal for identifying where known malicious behaviours interact with legitimate behaviours which – given current risk management capabilities – simply are not observable in existing cyber-risk management activities, and are as such not yet known.

1C **Identify candidate controls.** This identifies controls $C_1 - C_N$, and applies to managing both negative behaviours and protecting positive behaviours. Involving stakeholders will make this more tractable. Once conducted, assessments may be reusable, making it less demanding over time and akin to maintaining an ongoing *risk register*. Such a register would describe concerns to manage (left-side of Fig. 2), and a *behaviour register* of existing behaviours to preserve (right-side of Fig. 2). It may not be possible to confirm that all behaviours and associated controls in the system have been identified, but efforts to do so should be documented.

Step 2. Map connections between behaviours and system assets.

2A **Identify sociotechnical representations of behaviours.** For each Control C in $C_1 - C_N$, identify the Environment, or *cyberplace* [64]; the Behavioural determinants, Individual factors, and related

data representations as recorded in IT systems that it acts upon (such as registered users or roles, the systems they use for their work, etc.). User activities must translate to user or behaviour representations (data or rules, Fig. 1), or system elements, for a cyber-risk manager to be able to work directly with the information. Behaviour change approaches emphasise that it is critical to involve stakeholders in identifying target behaviours.

2B **Map behaviour determinants to technical features.** This will relate the impacts of controls on Environment and Behaviour to the Individual. For specific behaviours and their candidate controls, map data and systems to COM-B properties [77]. This can, for instance, map Capabilities to rules for permitted activity, or account properties; map Opportunity to restrictions on account access (such as registration requirements, or rules for signalling malicious behaviour); map Motivation to assumptions about workload/effort around what users will need to do to have access to a service (including technical knowledge). Having an Opportunity facilitated in technology does not necessarily mean that it is easily accessible. For instance, target-hardening efforts may make a system less accessible to legitimate users. For this reason, a user having access to – and being present in – an IT environment should be managed as a conscious Control decision.

Step 3. Address linkages between negative and positive behaviours and/or controls. Controls are engineered mechanisms [57] – it may be assumed but is not always assured that a control precisely addresses only the entity or activity it is intended to act upon. This means there is scope to address linkages. Controls and Behaviours must both be assessed together in an iterative manner. If it is found that any mapping of COM-B features to user, activity, or system entities overlaps between the *negative* and *positive* sets, it should be assumed that there is a legitimate group of users which will be affected by a cyber-risk management control if it is deployed. For instance, specific access restrictions may be activated by particular device or account details, but these rules might affect legitimate users sharing the same traits. Linkages would require *remediation* (see Section 4.4) to break, or record and compensate for, the shared dependency between positive and negative behaviours. The number of linkages is a basic indicator of potential harms and a lack of precision in the candidate control. A focus on the elements of the right-hand side would traditionally be associated with security usability, and the left-hand side associated with (cyber)crime reduction. System managers must decide how much *interference* they are willing to manage themselves, or unwittingly place upon legitimate users as a problem for them to potentially fix themselves after deployment. Efforts in national policy to reduce the burden of password-based authentication on users, and combine this with background technical controls (as in UK policy, Section 2.5) is an example of a nuanced approach which aims to prevent malicious activities *while also* reducing the harms placed upon users to have legitimate access.

4.4. Managing for the precision of risk controls

If a control affects both positive and negative behaviours, there may be a need to *reconsider* it. This would involve searching for a candidate Control which does not act on shared determinants, but only on negative behaviour determinants. With adaptation, current risk management processes would accommodate this, including searching for existing solutions already available to the risk owner. This highlights the need to take a mechanistic approach to understanding the role of security-related technologies in real-world systems [98]. Precise approaches for achieving this must be developed, where existing risk management guidelines can be adapted to identify controls which appropriately address a risk, relative to other activities already active in the system.

If a Control is adaptable, it can be *refined* – this applies more so to Controls which can be configured in how they interact with People, such as detection rules for system/online behaviour. We make an assumption that cybersecurity controls are generally deployed without an initial check of whether they carry the kind of *residual risk* which can result in *unintended harms*. There must be agreement with stakeholders that a control adequately minimises or avoids harms. If there is an expectation of potential harms to legitimate users – where negative and positive behaviour determinants interact – there may be a choice to *compensate* for the harm, and accept a candidate control but with additional compensatory measures. This may happen if a control is deemed necessary but expected to be short-lived (such as to address an emergent security threat). Refinements may be realised through e.g., configuration of data processing rules, policies for user identification and verification, user behaviour detection rules, and device detection and management rules.

Any lack of knowledge or expectations around the knock-on effects of a cybersecurity control should be logged as a residual risk ('unidentified risk' as in 27005:2011 [66]). This may be the case if a control is relatively novel. This relates to *ongoing attentiveness* [79] to making systems work together (Section 6), and not assuming deployment of a control as a final measure of its success in a sociotechnical system. This is realised most readily by measuring the performance of the system. The process should include input from non-security stakeholders, where their perception of consequences of cyber-risks must be considered [5]. Existing risk management approaches already advocate this, but not necessarily the residual risk of controls for legitimate users, or how to identify this particular kind of risk.

Application of the concepts in the framework leads to consideration of the connections between malicious activities to prevent, and legitimate behaviours to preserve. It would be regarded as increasingly common that two-factor authentication (2FA) technologies could be used to provide improved security. However, if the second factor is a phone-based one-time code, it presumes all users have a smartphone and are comfortable using it for the service. This may not be the case if a personal smartphone is expected to be used for an employer's 2FA scheme, for instance.

4.5. Measuring the precision of risk controls

It is at present difficult to immediately describe how to measure the precision of cyber-risk controls, as this is a challenge not already instrumented in managed systems. Also, as mentioned in Section 2.1, users and behaviours adversely affected by blunt controls in a persistent way may be comparatively few. They may otherwise be removed or blocked from the observable system (and so not be visible).

It is then necessary in the first instance to consider approaches which have potential to refine the search for imprecise cyber-risk controls. This immediately raises a need to see (i) what processes exist which are outside of managed systems but tell us something about activity within them, and (ii) where measures internal to a system are capable of capturing potentially subtle imprecision and unintended harms.

Where precision is considered in protecting users and deferring malicious activity, one approach is by understanding examples of fine-tuned precision in the system, for example making high-end earphones which are moulded to a specific person's ears, reducing their 'stealability' [41]. We propose something of a similar nature here, in terms of crafting a response to understand activities in the system more easily as those of a legitimate user. This would focus more on the existing behaviour of the user that we wish to preserve, thereby reducing impacts upon the determinants of their existing behaviour when targeting malicious activity 'around them' elsewhere in the system.

One measure of precision is then the lack of distinguishing features observable between negative and positive behaviours in the same space. Considering crime reduction (Section 2.2) and behaviour

preservation techniques (Section 2.3), this would be to say that these parallel efforts should be targeting different determinants. Hence we have difficulties in online spaces, for instance, where malicious entities can pretend to be legitimate ones, and also where malicious entities have just as much access to the same tools used for legitimate behaviours (such as consumer software and hardware, etc.). It is then a measure of imprecision created by the lack of understanding of user behaviour, which would be addressed by having a view of the behaviours that contribute to protected digital assets (Section 3.3). The Lockheed-Martin 'cyber kill chain' classifies intruder activities as being different kinds of malicious behaviours leading to exfiltration of data [31]. To support precision in preventing and preserving different behaviours, a comparable solution, would be to have knowledge in the system about behaviour classifications which are linked to an activity representation within a managed data asset that is being protected. Put another way, the killchain concept identifies the 'footprint' of negative behaviours as seen within IT systems – we argue that there is no complementary approach as yet for recording footprints of positive behaviours.

There are examples of unintended harms – lack of precision – which can be observed *outside* of a system. One example is capturing user experiences and sentiment. This can include app marketplace reviews (for specific apps, and especially where an app is being promoted as the main way to reach a service provider) – the security team could liaise with teams responsible for communications or troubleshooting, for instance. Another example is consumer groups, such as Which? in the UK, or specific dedicated examples such as Mozilla's 'Privacy Not Included' [80] guidance for consumer IoT purchases. Another source is security and/or technology journalists such as The Register, Bruce Schneier or Brian Krebs – specifically here, it is not just that they may write about issues that can come to a service manager's attention through publicity, but also that as public figures in security they may be conduits of user concerns (and amplifiers of unintended consequences experienced by a potentially relatively small proportion of users). These approaches are important, because imprecise controls may remove some legitimate users' *opportunity* or *capability* to enact a behaviour within a system; these affected users then cannot bring attention to their difficulties from within the system itself.

Critically, reports of user experience and sentiment will rely on a certain amount of power and self-determination to be in the system being exercised by individuals in that system. This may not always be the case – interventions can in some cases be intended to remove individual users from a difficult situation that they do not know they are in. This may be the case in situations of tech-abuse [1], for example. This is to say that the risk owner cannot solely rely on legitimate users to be aware of and raise concerns for themselves. The crime reduction concept of manager as guardian of legitimate users is then critical, and must not be forgotten [60] in the pursuit of *target hardening*. This is especially the case for user groups who have experienced long-term prevention of access, or who do not recognise a negative experience as not being the intended experience. An example would be if users are not able to access particular kinds of support through government services online, if they have had difficulties accessing comparative in-person services previously, they would believe their *capability* to be low, but this also impacts *motivation*. These factors then suggest a need to proactively seek out the intended users of online services before they experience problems reaching those services, and not just online. This informs what the role would be of a *cyber-risk guardian*. A simple example would be online security advice for citizens, which if only available on a website would require them to go online and actively search for it – instead, those giving the advice ought to seek out citizens offline, to bring the advice to those who would benefit from it.

Proactive efforts by a cyber-risk owner to measure precision of controls can also happen on their own technical platform. This can include systems logs where an example would be users leaving a platform

or accessing it less. There are inherent privacy issues with profiling user behaviour, but if measures such as these are being collected for 'system performance' purposes already, they may also be used to signal potential harms. At a user community level, the utility of platforms means that individuals may move to another platform because people they know cannot access the platform in question. However, in organizations, for instance, the burden of additional costs is still something of a *hidden cost* in the system. Users – in this case employees – may put in extra time for security beyond their normal work, so extended work hours could also mask difficulties in accessing systems.

From a design perspective, troubleshooting problems is an activity which arguably happens too late to avoid some harms which cannot be 'reversed'. Looking instead to the design of a system, security personas can help [49]. By developing representative user personas, it may be possible to identify who the cyber-risk owner believes malicious and legitimate users are, in terms of the kinds of Factors discussed within our framework. This, however, requires risk owners to be included in the design or renewal of systems, which at present cannot be assumed. If this is possible, measurement becomes less about finding a lack of precision, but more about maintaining a designed-in precision. This would be a precision which clearly separates determinants of malicious and legitimate behaviours believed to be possible in the managed system, updating these based on changes to the threat environment but also the climate of use. Examples of the latter would be the addition of new features to a service platform, or changes to the workforce in a company due to a business merger or introduction of a new business application. These would introduce new behaviours, akin to 'changing' behaviour by introducing previously unknown behaviours, which then must be preserved by the cyber-risk guardian.

5. Use cases in existing environments

To aid with the illustration of the steps, we will discuss the framework in the context of two real-world settings: (a) organization and (b) social media platforms (*SMPs*).

5.1. Use Case 1 – Phishing prevention in organizations

Step 1. Record behaviours in the system

1A A company may be instructing staff not to 'click on links' in emails, with the broad view that they are the *most likely* attack vector for phishing or targeted malware attacks.

1B Staff in the company may be legitimately sending emails to colleagues, or receiving emails from collaborators outside of the organization, which include URL links to legitimate, harmless online resources.

1C A question then in Step 1C is whether to instruct staff to never click on links in emails, and instead make it possible to complete work without ever needing to link to online resources. Another candidate control would be to purchase and provide two-factor authentication (2FA) tokens for staff, as in the Google example discussed earlier [72]; this would reduce any harms created from inadvertently sharing credentials if duped by a phishing attack.

Step 2. Mapping Connections between behaviours and System Assets

2A Ideally the company would catalog which business activities rely on URL links in emails, how many emails staff receive, etc. This is often not done at present, so the scale of the impact of prohibiting sharing of URLs is often not understood.

2B If users cannot share resources through URLs, they may employ other more costly – and potentially more dangerous – means. For instance, downloading content directly from URLs and sharing this by email instead; this inadvertently re-purposes corporate email as a data-sharing platform. Ideally then a corporate file-sharing application would be introduced to reduce employees' perception of the need to use email for this purpose. Such a solution would also need to be usable and accessible (Step 3).

Step 3. Identifying Linkages between Negative and Positive behaviours and Controls

Issues of control precision may arise, for instance, when differentiating between legitimate activities within a company, and with others outside of the company. A file-sharing platform as a solution for sharing resources may become relatively straightforward to use *within* a company, and using internal resources. It may become excessively difficult to use when collaborating with others who are not in the same company [16], inadvertently encouraging employees to make local, unsecured copies of important files to be able to share them with others outside of the company. This would potentially prompt *refinement* of the control, to provide a solution to avoid sharing of URLs inside the organization, and increased technical monitoring and promotion of reporting points for staff within the organization with regards to suspicious emails.

If controls focus on how employees treat URLs in emails, some staff may find it more difficult than others to identify particular kinds of malicious email [99]. A reporting point for malicious emails can then be valuable, or support for when staff are uncertain. This task could be seen as too difficult to enact on a regular basis (based on, for instance, how many emails an employee receives with links in them). In this case, the token-based solution mentioned earlier, or a similar 2FA approach, may be employed for specific groups of employees. For instance, outward-facing teams such as invoicing, recruitment, or Public Relations (PR) teams may receive many emails with links or requests for payment, with relatively little or no pre-existing shared context with the sender.

5.2. Use Case 2 – Social media platforms

The domain of Social Media Platforms (SMPs) is one within which platform operators have needed to perform multiple activities, including: (a) iterate controls for security and privacy, (b) to ensure confident use by a range of different legitimate users, and (c) identifying and preventing malicious and negative activity. The proposed framework is applied to highlight gaps in existing risk-management approaches.

Step 1. Record behaviours in the system

1A Online abuse continues to be an issue as technology and the Internet are interleaved with our daily routines. Online abuse includes behaviours such as trolling, online harassment, stalking, bullying, and online threats [56,61,101]. The increased use of SMPs like Facebook and Twitter allow for continuous contact between offenders and targets without regard for physical and temporal distance [61]. This constitutes negative behaviour to be identified and prevented on SMPs.

1B Simultaneously, use of SMPs continue to grow among teenagers and adults [22] and is encouraged for their beneficial effects [45]. In this instance, there are two positive outcomes to be preserved: encouraging continued use of SMPs, while also lowering users' risks of becoming targets as they converge with offenders in the same online social space.

1C To address the negative behaviour, SMPs introduce controls to minimise its occurrence and impacts on users (e.g., [46,96,102]). There is the use of privacy settings and controls that allow account holders to manage accessibility to content via blocking or filtering [46,96,102]. Facebook later

introduced the "friend list" feature to dictate the types of content each list has access to [47]. Snapchat provides finer granularity in controls, such as "Who can view my Story" and "Who can contact me" [96].

Another type of control is the introduction of clear community rules. The Snapchat community guidelines explicitly prohibit harassment, bullying, impersonation or violence, and encourage account holders to report these behaviours [97]. SMPs listed punishments of different severity in guidelines, from the removal of content, to termination of an account, to the possibility of activity being reported to law enforcement agencies [97]. In some cases, the platforms attempt to include other stakeholders in their controls. Snapchat encourages parents to help adolescents in managing their accounts [96]. Parents have also advised their children to manage privacy by providing false information [34].

Personal privacy controls are realised in part through security controls which maintain a safe environment which users can trust. A user may exercise a privacy decision through a service – security controls can serve to create the environment which enforces those decisions. An example would be ensuring that an SMP user cannot be reached by another user who they have blocked or not explicitly provided visibility to.

Step 2. Mapping Connections between Behaviours and System Assets

2A **Negative Behaviours:** Current literature establishes a range of factors contributing to the rise of online abusive behaviours. Factors to consider at the individual level include pro-victim attitudes [44], perceptions of norms and injustice [17], and the contexts of exchanges [17,105]. Other relevant factors of cyberbullying relate to features of cyberspace, such as the anonymity and distance between users which can result in a sense of impunity and deindividuation. This can lead to adoption of online aggressive behaviours [56,59,93]. The nature of online media also means that users are removed from direct confrontation or consequence for their own behaviours [59,93]. Another feature is the scalability of the Internet, which allows multiple individuals to participate simultaneously in bullying behaviours [59].

Another factor is the possible overlap between offenders and targets in cyberbullying and cyber-interpersonal violence [24,105]. This overlap is exacerbated by a reliance on users to be proactive. In addition, poorly managed privacy controls on accessibility of social network content are linked with an increased probability of becoming a victim and offender of cyber-interpersonal violence [24].

Controls: One factor affecting the utilisation of controls is the 'privacy paradox', where there is a disparity between expressed privacy concerns and privacy-related behaviours [8]. For instance, users have reported utilizing features such as friends-only content accessibility, but at the same time accepting large numbers of friend requests from individuals who may not be seen as friends beyond the context of the SMP [35].

2B There is some evidence supporting the effectiveness of SMP controls. Younger users of SMPs tend to be more proactive in adopting existing accessibility controls and settings [4,8,34,35,71]. A comprehensive review on cyberbullying also found that blocking cyberbullies is among the most common strategies used and recommended among children and adolescents [4,55].

Other factors affecting the effectiveness of existing controls, especially privacy controls, are users' engagement, proactivity toward privacy, and technical skills [8,10]. These must be balanced with users' aims to communicate with others, potentially opportunistically or openly. This points to a combination of COM-B elements [77].

This can require approachable means for finding other users on the same SMP, reaching others with messages they potentially were not expecting, and being able to tune interests to define the messages which are received from other accounts. In terms of security and privacy, this would require a blend of controls to prevent negative behaviours and realise user intentions.

Step 3. Identifying Linkages between Negative and Positive behaviours and Controls

Existing controls on SMPs address different COM-B characteristics that affect both behaviours that we wish to preserve (use of trustworthy SMPs) and prevent (online abuse). First, there is an inherent source of interference in the nature of the environment and users' motivations. The primary purposes for using SMPs include expressing one's identity digitally, maintaining and enhancing existing offline and online relationships, and creating new social relationships [109]. To reach their goals, both legitimate and malicious users share some degree of information such as names and email addresses [100,109]. These requirements, along with the small to moderate effects between privacy concerns and users' utilization of privacy controls [8,10,35,71], suggest increased opportunity for malicious behaviours as existing controls do not fully align with legitimate behaviours.

The second source of interference is the tension that stems from differences in the dynamics of online and offline social relationships. Online SMPs tend to oversimplify social ties into friends and not-friends [109]. Such dichotomous definitions do not always reflect the fluidity of social relationships in the offline world, adding to the effort required to maintain online privacy. In addition, users of online SMPs assign different values to different types of personal and sensitive information in cyberspace [2,4,71,100]. Variations in value assignment can interfere with perceptions of risks and opportunities, in turn affecting users' utilization of existing controls.

6. Discussion

Our framework combines existing capabilities across disciplines, highlighting where adjustments can better manage sociotechnical risks. An existing risk register can be extended to log existing positive behaviours, but this may require concerted effort and knowledge of activities in the system which have positive effects. Communication is required with specific stakeholders such as Human Resources departments, user advocacy groups, etc. This is more tractable than determining where users have been 'forgotten' or removed by harmful risk controls [25].

A risk owner may not be willing – or able – to *reconsider* or *refine* a control (Section 4.4). At an extreme, they may act to remain ignorant of potential harms created by a cybersecurity control, as 'organised irresponsibility' [5]. This introduces its own risk, of assuming that a control will not have impacts for legitimate users or that impacts transferred to users are trivial, which undermines security assurances. This would be a form of *risk acceptance*, which in light of unintended harms would be *imposed acceptance* on users (as risk dumping [25]).

We propose an approach to risk management which combines prevention and preservation of behaviours to avoid linkages between them. This would bolster what Molotch [79] advocates in safety management, to *"add to rather than subtract from our well-being"*, by providing secure IT environments which are accessible to intended users. Molotch also advocates *ongoing attentiveness* to the management of risks, which in this context would be regular oversight and dialogue with stakeholders. At present, security guidelines signpost seemingly few points at which to engage parties with localised knowledge of user needs, or points at which to gather knowledge of legitimate uses of a managed service/system in order to better understand what is to be protected.

6.1. Limitations to the proposed approach

There are limitations to the application of the proposed framework, and what it can achieve. Similar to the work described in [25] for identifying unintended harms of cybersecurity controls, the outcomes of applying our framework are generative. The framework guides exploration of potential unanticipated side-effects of cyber-risk controls. The generation of potential harms to investigate is here grounded in technological artefacts familiar to the system security manager, as a starting point to reach beyond the security function and engage with the wider service/environment ecosystem. This means that the range of potential harms being explored is not assured to be complete; as noted in the consideration of how to manage the precision of cyber-risk controls (Section 4.4). Any exploration conducted ahead of deployment can spare legitimate users from unintended burden. Leaving the resourcing and design of 'compensation' measures until the point of deployment (or after) is arguably too late, so the more this unwanted outcome can be reduced, the more accessible systems will be for those users anticipated to use them for legitimate activities.

The framework would be more difficult to use if there is not either an ongoing 'behaviour register' or ongoing engagement with system/service users whenever a control is deployed (as described in Section 4.5). A behaviour register assumes that the cyber-risk manager has sight of the state of the data assets they are tasked with protecting, i.e., they can monitor their properties in the process of determining if those assets are secure.

Where a behaviour register is lacking, the steps for stakeholder modelling as outlined by Poller et al. [87] serve as an existing approach which can be adapted, driven by stakeholder interactions, but with the effect of assessing security controls rather than the management of perceived threats. This includes the following: identify stakeholder collaborations; identify collaborations which may be impacted by deploying the control (including those perceived to be protected by the control), and; identify 'transmission factors' across stakeholder inter-dependencies (which here would be harms imposed upon legitimate behaviours [25] by security management efforts). Alternatively, a preliminary version of applying the framework in an ecosystem would be to have a first conversation about controls, and how the system is being used – this is in the spirit of the line of work started with the 'Users are not the Enemy' research [3], through to the Security Dialogues approach [7].

The framework itself will not make controls more precise – it serves as a structure to identify imprecision in existing controls. This in turn relies on having an understanding of both 'benign', legitimate behaviours in the system, and how those may be affected by an existing or candidate security control. We simplify to control features as these can be discussed in terms of how they can affect aspects of behaviours. In security, how security controls interact with the pillars of the Capability-Opportunity-Motivation (COM-B) model [77] is a growing body of research in itself; here we focus consideration of COM pillars between a legitimate behaviour and a security control. This activity may require enumerating over individual controls against all (known) behaviours.

Using the COM-B pillars loses some of the nuance of those pillars when they are applied in practice. For instance, crime reduction techniques can include replacing potential paths to malicious behaviour with paths toward positive behaviours. The simplification to COM-B pillars is to serve as a *trading zone* [98], a shared language between the security function and other stakeholders, where a dialogue between representatives of all affected parties needs to happen; to not accommodate such a dialogue is to assume that cyber-risk controls never have the potential to impact legitimate users. Such an assumption creates risks of its own if unintended harms to legitimate users are increasingly identified only during active deployment of controls (at which stage, the capacity to prevent or undo harms is vastly reduced).

6.2. Future directions

Future work will also explore how existing cyber-risk management standards can be adapted and extended to promote precision in sociotechnical risk management. This would include how cyber-risk management can be made a more inclusive process, in a sustainable way; recording what *should* be in a complex IT system – in a sociotechnical register that includes both a 'behaviour register' and a more familiar risk register – is potentially a much bigger task than recording what *should not* (the risk register alone). Security managers should also be supported in considering system activity that has traditionally been outside of their domain; tool support is another avenue.

6.3. Recommendations

We make a number of recommendations for moving toward more precise sociotechnical cyber-risk management, including:

- **Extend the management of digital assets to include user *activity representations***. In cyber-risk management processes, we must go beyond only considering data and components involved in activities within risk registers. There is a need to include representations of active user behaviour, the behaviour which when enacted, *creates and changes* the data asset to produce its perceived value. OCTAVE Allegro [21] advocates similar initiatives. There is a need for methods for engagement with other stakeholders, such as: premortems to help understand how services might fail intended users in the future [69] (and any role of security in this). Existing analysis approaches can be useful for this purpose; 'security fictions' which bring undetected security issues into view for security specialists [76]; security personas to represent user requirements in a designed system [49], and; lightweight risk analysis methods to capture the context of use [52]. Such approaches can be used to discuss cyber-risk management policies when they are written or reviewed, to ensure that rules are reasonable and will not create unintended harms themselves. Risk management guidelines are gradually developing further capabilities to respect legitimate system use when managing malicious activity, as evidenced by OCTAVE Allegro; in time there should be a shift in cyber-risk management policies to include more sociotechnical elements, linked explicitly to traditionally technical elements, to surface how they affect each other.
- **Develop control portfolios to accommodate precision.** There must be capacity to tailor controls to match specific negative behaviour controls, and leave positive behaviours alone. The work of Hatleback and Spring [57], and Chua et al. [25], provide a basis for terminology to navigate between prevention and preservation of behaviours. There is then a link between management of behaviours in a system and the availability of solutions which can be used to realise the aims of that management activity. Identifying where solutions are not precise enough is the first step, but there must then be some capacity to replace or configure existing controls to realise a more appropriate solution. An example would be the emergence of password managers to replace the burden of users managing a huge number of knowledge-based credentials; password managers may or may not be available to users within specific managed environments. The importance of innovation for creating workable solutions is highlighted in both crime prevention (e.g., Vibrant Secure Function Framework (VSFF) [107]) and behaviour change (e.g., the Behaviour Change Ball [58]). Referring to regulations which govern how a system should be managed, there are examples of the maintenance of adaptive guidelines, for instance to govern Internet-of-Things technologies [19], just as there are co-evolutionary approaches to crime prevention [65]. Where cybersecurity policies may

be updated in light of new threats, they should equally be reviewed when new solutions become available. This does already happen to some extent, but what is missing is the foresight to ensure that 'less harmful' solutions are available when policies are reviewed. There may be risk managers who know that their IT systems are not perfect for legitimate users, but who lack 'better' options to include in their policies.

- **Measure control precision**, and with this an understanding of how unintended impacts upon legitimate users manifest in a system. We present a simple measure, of the number of overlapping factors between negative and positive behaviours. For instance, legitimate activities and phishing attacks both use hyperlinks. One observed reason for behaviour management activities failing is a lack of consideration for subgroups [83] – an activity to manage behaviour can succeed, not have any effect, or potentially backfire at the same time, as different user groups experience (or fail to experience) the same outcome. Another approach that is proposed is then causal analysis related to relevant factors, where our framework prompts risk owners to identify these factors in cooperation with other stakeholders. Failures of behaviour management can then be utilised as a resource from which to learn how to design future interventions. A capacity to measure performance in *guarding* user activities could then act as a metric for signalling when policies need to be reviewed.

7. Conclusion

We describe a framework for management of the concurrent prevention and promotion of different security and privacy behaviours in a managed IT environment. This framework leverages risk management approaches familiar to practitioners, and a synergy of approaches from (cyber)crime reduction and behaviour change science. The definition of digital assets for risk management must explicitly include representations of user behaviour in managed systems; the role of stakeholders and how to engage with them is underspecified in cyber-risk management standards, and; more must be done to measure unintended harms upon legitimate users, and develop candidate cyber-risk controls with a precision that avoids impacts on determinants of protected user behaviours. Future work would explore the notion of sociotechnical precision in cybersecurity and cyber-risk management, with a real-world environment, related stakeholders, and discernible vulnerable populations.

Acknowledgments

We thank Paul Ekblom for comments on an earlier version of this work presented at the 10th Workshop on Socio-Technical Aspects in Security (STAST 2020), as well as attendees of STAST/ESORICS 2020 who provided comments on this work.

References

[1] Tech vs. Abuse, Tech vs. Abuse 2.0 – Design Principles, 2019, (Accessed 06/02/2021). https://www.techvsabuse.info/design-principles.

[2] A. Acquisti and R. Gross, Imagined communities: Awareness, information sharing, and privacy on the Facebook, in: *International Workshop on Privacy Enhancing Technologies*, Springer, 2006, pp. 36–58. doi:10.1007/11957454_3.

[3] A. Adams and M.A. Sasse, Users are not the enemy, *Communications of the ACM* **42**(12) (1999), 40–46. doi:10.1145/322796.322806.

[4] M. Adorjan and R. Ricciardelli, A new privacy paradox? Youth agentic practices of privacy management despite "nothing to hide" online, *Canadian Review of Sociology* **56**(1) (2019), 8–29. doi:10.1111/cars.12227.
[5] I. Agrafiotis, M. Bada, P. Cornish, S. Creese, M. Goldsmith, E. Ignatuschtschenko, T. Roberts and D.M. Upton, Cyber harm: Concepts, taxonomy and measurement, *Saïd Business School WP* **23** (2016).
[6] C. Alberts, S. Behrens, R. Pethia and W. Wilson, Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) Framework, Version 1.0, Technical Report, CMU/SEI-99-TR-017, Software Engineering Institute, Carnegie Mellon University, 1999.
[7] D. Ashenden and D. Lawrence, Security dialogues: Building better relationships between security and business, *IEEE Security & Privacy* **14**(3) (2016), 82–87. doi:10.1109/MSP.2016.57.
[8] S.B. Barnes, A privacy paradox: Social networking in the United States, *First Monday* **11**(9) (2006).
[9] L.K. Bartholomew, G.S. Parcel and G. Kok, Intervention mapping: A process for developing theory and evidence-based health education programs, *Health Education & Behavior* **25**(5) (1998), 545–563. doi:10.1177/109019819802500502.
[10] L. Baruh, E. Secinti and Z. Cemalcilar, Online privacy concerns and privacy management: A meta-analytical review, *Journal of Communication* **67**(1) (2017), 26–53. doi:10.1111/jcom.12276.
[11] BBC, Google executive warns of face ID bias, 2018, Accessed: 27th Feb 2021, https://www.bbc.com/news/technology-44977366.
[12] BBC, Overtly sexual, 2021, cow blocked as Facebook ad, 2021, Accessed: 27th, https://www.bbc.com/news/technology-55981602.
[13] BBC, AWS: Amazon web outage breaks vacuums and doorbells, 2020, Accessed: 27th Feb 2021, https://www.bbc.com/news/technology-55087054.
[14] BBC, WhatsApp extends 'confusing' update deadline, 2021, Accessed: 27th Feb 2021, https://www.bbc.com/news/technology-55683745.
[15] BBC, Last of Us Part II: Is this the most accessible game ever? 2020, Accessed: 27th Feb 2021, https://www.bbc.com/news/technology-53093613.
[16] A. Beautement, M.A. Sasse and M. Wonham, The compliance budget: Managing security behaviour in organisations, in: *2008 New Security Paradigms Workshop (NSPW'08)*, 2008, pp. 47–58. doi:10.1145/1595676.1595684.
[17] L. Blackwell, T. Chen, S. Schoenebeck and C. Lampe, When online harassment is perceived as justified, in: *Twelfth International AAAI Conference on Web and Social*, Media, 2018.
[18] A.M. Bossler and G.W. Burruss, The general theory of crime and computer hacking: Low self-control hackers? in: *Cyber Crime: Concepts, Methodologies, Tools and Applications*, IGI Global, 2012, pp. 1499–1527. doi:10.4018/978-1-61350-323-2.ch707.
[19] I. Brass and J.H. Sowell, Adaptive governance for the Internet of Things: Coping with emerging security risks, *Regulation & Governance* (2020).
[20] BS, ISO, *BS ISO 31000:2018 – Risk management – Guidelines*, BS ISO, 2018, RM/1. ISBN 978 0 580 88518 1.
[21] R. Caralli, J. Stevens, L. Young and W. Wilson, Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process, Technical Report, CMU/SEI-2007-TR-012, Software Engineering Institute, Carnegie Mellon University, 2007.
[22] Pew Research Center, Demographics of Social Media Users and Adopters in the United States, 2019, https://www.pewresearch.org/internet/fact-sheet/social-media/.
[23] K.-S. Choi, Computer crime victimization and integrated theory: An empirical assessment, *International Journal of Cyber Criminology* **2**(1) (2008).
[24] K.-S. Choi and J.R. Lee, Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory, *Computers in Human Behavior* **73** (2017), 394–402. doi:10.1016/j.chb.2017.03.061.
[25] Y.T. Chua, S. Parkin, M. Edwards, D. Oliveira, S. Schiffner, G. Tyson and A. Hutchings, Identifying unintended harms of cybersecurity countermeasures, in: *2019 APWG Symposium on Electronic Crime Research (eCrime)*, IEEE, 2019, pp. 1–15.
[26] R.V. Clarke, Situational crime prevention: Its theoretical basis and practical scope, *Crime and Justice* **4** (1983), 225–256. doi:10.1086/449090.
[27] R.V. Clarke, *Situational Crime Prevention: Successful Case Studies*, Harrow and Heston Publishers, Albany, NY, 1997.
[28] J. Clear, *Atomic Habits: An Easy & Proven Way to Build Good Habits & Break Bad Ones*, Penguin, 2018.
[29] L.E. Cohen and M. Felson, Social change and crime rate trends: A routine activity approach, *American Sociological Review* (1979), 588–608. doi:10.2307/2094589.
[30] J.D. Collins, V.A. Sainato and D.N. Khey, Organizational Data Breaches 2005-2010: Applying SCP to the Healthcare and Education Sectors, *International Journal of Cyber Criminology* **5**(1) (2011).
[31] W. community, Cyber kill chain, 2021, (Accessed 06/02/2021). https://en.wikipedia.org/wiki/Kill_chain#The_cyber_kill_chain.
[32] D.B. Cornish and R.V. Clarke, Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention, *Crime Prevention Studies* **16** (2003), 41–96.

[33] L. Coventry, P. Briggs, D. Jeske and A. van Moorsel, SCENE: A structured means for creating and evaluating behavioral nudges in a cyber security environment, in: *International Conference of Design, User Experience, and Usability*, Springer, 2014, pp. 229–239.
[34] K. Davis and C. James, Tweens' conceptions of privacy online: Implications for educators, *Learning, Media and Technology* **38**(1) (2013), 4–25. doi:10.1080/17439884.2012.658404.
[35] B. Debatin, J.P. Lovejoy, A.-K. Horn and B.N. Hughes, Facebook and online privacy: Attitudes, behaviors, and unintended consequences, *Journal of Computer-Mediated Communication* **15**(1) (2009), 83–108. doi:10.1111/j.1083-6101.2009.01494.x.
[36] J.E. Eck and R.V. Clarke, Classifying common police problems: A routine activity approach, *Crime prevention studies* **16** (2003), 7–40.
[37] M. Edwards, G. Suarez-Tangil, C. Peersman, G. Stringhini, A. Rashid and M. Whitty, The geography of online dating fraud, in: *Workshop on Technology and Consumer Protection (ConPro)*, 2018.
[38] P. Ekblom, The security function framework, in: *Design Against Crime: Crime Proofing Everyday Products*, P. Ekblom, ed., Lynne Rienner Publishers, 2012, pp. 9–36, Chap. 2. doi:10.1515/9781588269409-005.
[39] P. Ekblom, Crime prevention through product design, in: *Handbook of Crime Prevention and Community Safety*, Taylor & Francis, Abingdon, 2017, pp. 207–233. doi:10.4324/9781315724393-10.
[40] P. Ekblom, Technology, opportunity, crime and crime prevention: Current and evolutionary perspectives, in: *Crime Prevention in the 21st Century*, Springer, 2017, pp. 319–343. doi:10.1007/978-3-319-27793-6_19.
[41] P. Ekblom et al., Happy returns: Ideas brought back from situational crime prevention's exploration of design against crime, 2011.
[42] P. Ekblom, K.J. Bowers, L. Gamman, A. Sidebottom, C. Thomas, A. Thorpe, M. Willcocks et al., Reducing bag theft in bars, Design Against Crime: Crime Proofing Everyday Products, *Crime Prevention Studies* **27** (2012).
[43] P. Ekblom and M. Gill, Rewriting the script: Cross-disciplinary exploration and conceptual consolidation of the procedural analysis of crime, *European Journal on Criminal Policy and Research* **22**(2) (2016), 319–339. doi:10.1007/s10610-015-9291-9.
[44] L.C. Elledge, A. Williford, A.J. Boulton, K.J. DePaolis, T.D. Little and C. Salmivalli, Individual and contextual predictors of cyberbullying: The influence of children's provictim attitudes and teachers' ability to intervene, *Journal of Youth and Adolescence* **42**(5) (2013), 698–710. doi:10.1007/s10964-013-9920-x.
[45] N.B. Ellison, C. Steinfield and C. Lampe, The benefits of Facebook "friends:" Social capital and college students' use of online social network sites, *Journal of Computer-Mediated Communication* **12**(4) (2007), 1143–1168. doi:10.1111/j.1083-6101.2007.00367.x.
[46] Facebook, Abuse Resources, 2020, Accessed: 10.09.2020, https://www.facebook.com/help/726709730764837/?helpref=hc_fnav.
[47] Facebook, Friend lists |Facebook Help Centre, 2020, Accessed: 08.12.2019, https://www.facebook.com/help/204604196335128.
[48] S. Fahl, Y. Acar, H. Perl and M. Smith, Why Eve and Mallory (also) love webmasters: A study on the root causes of SSL misconfigurations, in: *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, 2014, pp. 507–512. doi:10.1145/2590296.2590341.
[49] S. Faily and I. Fléchais, Barry is not the weakest link: Eliciting Secure System Requirements with Personas, 2010.
[50] M. Felson, Those who discourage crime, *Crime and place* **4** (1995), 53–66.
[51] M. Felson and L.E. Cohen, Human ecology and crime: A routine activity approach, *Human Ecology* **8**(4) (1980), 389–406. doi:10.1007/BF01561001.
[52] I. Flechais, M.A. Sasse and S.M. Hailes, Bringing security home: A process for developing secure and usable systems, in: *Proceedings of the 2003 Workshop on New Security Paradigms*, 2003, pp. 49–57. doi:10.1145/986655.986664.
[53] B.J. Fogg, *Tiny Habits: The Small Changes That Change Everything*, Houghton Mifflin Harcourt, 2019.
[54] L.W. Green, Toward cost-benefit evaluations of health education: Some concepts, methods, and examples, *Health Education Monographs* **2**(1_suppl) (1974), 34–64. doi:10.1177/10901981740020S106.
[55] M.P. Hamm, A.S. Newton, A. Chisholm, J. Shulhan, A. Milne, P. Sundar, H. Ennis, S.D. Scott and L. Hartling, Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies, *JAMA pediatrics* **169**(8) (2015), 770–777. doi:10.1001/jamapediatrics.2015.0944.
[56] C. Hardaker, Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions, *Journal of Politeness Research* **6**(2) (2010), 215–242. doi:10.1515/jplr.2010.011.
[57] E.N. Hatleback and J.M. Spring, A refinement to the general mechanistic account, *European Journal for Philosophy of Science* **9**(2) (2019), 19. doi:10.1007/s13194-018-0237-1.
[58] A.-M. Hendriks, M.W. Jansen, J.S. Gubbels, N.K. De Vries, T. Paulussen and S.P. Kremers, Proposing a conceptual framework for integrated local public health policy, applied to childhood obesity-the behavior change ball, *Implementation Science* **8**(1) (2013), 46. doi:10.1186/1748-5908-8-46.
[59] S. Hinduja and J. Patchin, *Cyberbullying: Identification, Prevention, & Response*, Cyberbullying Research Center, 2018.

[60] M.E. Hollis, M. Felson and B.C. Welsh, The capable guardian in routine activities theory: A theoretical and conceptual reappraisal, *Crime Prevention and Community Safety* **15**(1) (2013), 65–79. doi:10.1057/cpcs.2012.14.
[61] T.J. Holt and A.M. Bossler, An assessment of the current state of cybercrime scholarship, *Deviant Behavior* **35**(1) (2014), 20–40. doi:10.1080/01639625.2013.822209.
[62] IEC, ISO, 31010: 2009 Risk management – Risk assessment techniques (2009). doi:10.3403/30183975.
[63] IEC, ISO, *BS ISO/IEC 25010:2011 – Systems and software engineering. Systems and software quality requirements and evaluation (SQuaRE). System and software quality models*, IEC, ISO, 2011, IST/15. ISBN 978 0 580 70223 5.
[64] C.C. Ife, T. Davies, S.J. Murdoch and G. Stringhini, Bridging Information Security and Environmental Criminology Research to Better Mitigate Cybercrime, 2019, arXiv preprint arXiv:1910.06380.
[65] 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. doi:10.1007/978-981-15-1304-6_22.
[66] ISO, IEC, IEC 27005: 2011 (EN) Information technology – Security techniques – Information security risk management, *ISO/IEC* (2011).
[67] Joint Task Force, Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy, (Final Public Draft) (SP 800-37 Rev. 2), Technical Report, National Institute of Standards and Technology, 2018.
[68] I. Kirlappos, S. Parkin and M. Sasse, Learning from "Shadow Security": Why understanding non-compliant behaviors provides the basis for effective security, in: *Workshop on Usable Security and Privacy (USEC'14)*, 2014, pp. 1–10.
[69] G.A. Klein, *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*, MIT Press, 2011.
[70] B. Knowles and V.L. Hanson, The wisdom of older technology (non) users, *Communications of the ACM* **61**(3) (2018), 72–77. doi:10.1145/3179995.
[71] S. Kokolakis, Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon, *Computers & Security* **64** (2017), 122–134. doi:10.1016/j.cose.2015.07.002.
[72] B. Krebs, Google: Security Keys Neutralized Employee Phishing, 2018, Accessed: 13th July 2020, https://krebsonsecurity.com/2018/07/google-security-keys-neutralized-employee-phishing/.
[73] K. Krol, M.S. Rahman, S. Parkin, E. De Cristofaro and E. Vasserman, An exploratory study of user perceptions of payment methods in the UK and the US, in: *Proceedings of the 10th NDSS Workshop on Usable Security (USEC 2016)*, Internet Society, 2016.
[74] J.R. Lee and T.J. Holt, Assessing the factors associated with the detection of juvenile hacking behaviors, *Frontiers in Psychology* **11** (2020), 840. doi:10.3389/fpsyg.2020.00840.
[75] C.D. Marcum, G.E. Higgins, M.L. Ricketts and S.E. Wolfe, Hacking in high school: Cybercrime perpetration by juveniles, *Deviant Behavior* **35**(7) (2014), 581–591. doi:10.1080/01639625.2013.867721.
[76] N. Merrill, Security fictions: Bridging speculative design and computer security, in: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2020, pp. 1727–1735. doi:10.1145/3357236.3395451.
[77] S. Michie, L. Atkins and R. West, *The Behaviour Change Wheel, a Guide to Designing Interventions*, 1st edn, Silverback Publishing, Great Britain, 2014, pp. 1003–1010.
[78] A. M'manga, S. Faily, J. McAlaney and C. Williams, Folk risk analysis: Factors influencing security analysts' interpretation of risk, in: *Thirteenth Symposium on Usable Privacy and Security (SOUPS'17)*, 2017.
[79] H.L. Molotch, *Against Security: How We Go Wrong at Airports, Subways, and Other Sites of Ambiguous Danger*, Princeton University Press, 2014. ISBN 9781400852338.
[80] Mozilla, *privacy not included, 2021, (Accessed 06/02/2021). https://foundation.mozilla.org/en/privacynotincluded/.
[81] B.B.C. News, Google thwarts Baltimore ransomware fightback, 2019, Accessed: 15th September 2020., https://www.bbc.co.uk/news/technology-48380662.
[82] B.B.C. News, Katrina O'Hara murder: Coroner recommends phone access changes, 2020, Accessed: 13th July 2020, https://www.bbc.co.uk/news/uk-england-dorset-51557476.
[83] M. Osman, S. McLachlan, N. Fenton, M. Neil, R. Löfstedt and B. Meder, Learning from behavioural changes that fail, *Trends in Cognitive Sciences* (2020).
[84] S. Parkin and K. Krol, Appropriation of security technologies in the workplace, Workshop on Experiences of Technology Appropriation: Unanticipated Users, *Usage, Circumstances, and Design* (2015).
[85] 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: *New Security Paradigms Workshop (NSPW'19)*, ACM, 2019.
[86] 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: *Workshop on Usable Security and Privacy (USEC'19)*, Internet Society, 2019.

[87] A. Poller, S. Türpe and K. Kinder-Kurlanda, An asset to security modeling? Analyzing stakeholder collaborations instead of threats to assets, in: *Proceedings of the 2014 New Security Paradigms Workshop*, 2014, pp. 69–82. doi:10.1145/2683467.2683474.
[88] J. Reason, Human error: Models and management, *Bmj* **320**(7237) (2000), 768–770. doi:10.1136/bmj.320.7237.768.
[89] K. Renaud, Accessible cyber security: The next frontier? in: *ICISSP*, 2021, pp. 9–18.
[90] K. Renaud and M. Warkentin, Using intervention mapping to breach the cyber-defense deficit, in: *12th Annual Symposium on Information Assurance (ASIA'17) June*, 2017, pp. 7–8.
[91] B.W. Reyns, A situational crime prevention approach to cyberstalking victimization: Preventive tactics for Internet users and online place managers, *Crime Prevention and Community Safety* **12**(2) (2010), 99–118. doi:10.1057/cpcs.2009.22.
[92] L. Romero, Experts say echo chambers from apps like Parler and Gab contributed to attack on Capitol, 2021, Accessed: 27th Feb 2021, https://abcnews.go.com/US/experts-echo-chambers-apps-parler-gab-contributed-attack/story?id=75141014.
[93] R. Sambaraju and C. McVittie, Examining abuse in online media, *Social and Personality Psychology Compass* **14**(3) (2020), 12521. doi:10.1111/spc3.12521.
[94] M.A. Sasse and I. Flechais, Usable security: Why do we need it? How do we get it? 2005, O'Reilly.
[95] R.G. Smith, N. Wolanin and G. Worthington, Trends & Issues in Crime and Criminal Justice No. 243: e-Crime solutions and crime displacement, 2003.
[96] Snapchat, Privacy Settings, 2020, Accessed: 07.03.2020, https://support.snapchat.com/en-GB/article/privacy-settings2.
[97] Snapchat, Community Guidelines, 2020, https://www.snap.com/en-US/community-guidelines.
[98] J.M. Spring, T. Moore and D. Pym, *Practicing a Science of Security: A Philosophy of Science Perspective, in: 2017 New Security Paradigms Workshop (NSPW'17)*, ACM, 2017.
[99] M.P. Steves, K.K. Greene, M.F. Theofanos et al., A phish scale: rating human phishing message detection difficulty, 2019, in: Workshop on usable security (USEC).
[100] M. Taddicken, The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure, *Journal of Computer-Mediated Communication* **19**(2) (2014), 248–273. doi:10.1111/jcc4.12052.
[101] The Crown Prosecution Service, Cyber/online crime, 2020, Accessed: 07.03.2020, https://www.cps.gov.uk/cyber-online-crime.
[102] TikTok, Safety Center, 2020, Accessed: 07.03.2020, https://www.tiktok.com/safety/resources/anti-bully?lang=en.
[103] J. Turland, L. Coventry, D. Jeske, P. Briggs and A. van Moorsel, Nudging towards security: Developing an application for wireless network selection for Android phones, in: *2015 British HCI Conference*, 2015, pp. 193–201. doi:10.1145/2783446.2783588.
[104] Y. Wang, Inclusive security and privacy, *IEEE Security & Privacy* **16**(4) (2018), 82–87. doi:10.1109/MSP.2018.3111237.
[105] E. Whittaker and R.M. Kowalski, Cyberbullying via social media, *Journal of School Violence* **14**(1) (2015), 11–29. doi:10.1080/15388220.2014.949377.
[106] M.T. Whitty and T. Buchanan, The online romance scam: A serious cybercrime, *CyberPsychology, Behavior, and Social Networking* **15**(3) (2012), 181–183. doi:10.1089/cyber.2011.0352.
[107] M. Willcocks, P. Ekblom and A. Thorpe, Less crime, more vibrancy, by design, Rebuilding Crime Prevention Through Environmental Design: Strengthening the Links with, *Crime Science* (2019), 216.
[108] R. Wortley, A classification of techniques for controlling situational precipitators of crime, *Security Journal* **14**(4) (2001), 63–82. doi:10.1057/palgrave.sj.8340098.
[109] C. Zhang, J. Sun, X. Zhu and Y. Fang, Privacy and security for online social networks: Challenges and opportunities, *IEEE network* **24**(4) (2010), 13–18. doi:10.1109/MNET.2010.5510913.
[110] V. Zimmermann and K. Renaud, Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset, *International Journal of Human-Computer Studies* **131** (2019), 169–187. doi:10.1016/j.ijhcs.2019.05.005.