CREST Centre for Research and Evidence on Security Threats

Simulated Phishing and Employee Cybersecurity Behaviour (SPEC)

FULL REPORT

SEPTEMBER 2020

Dr. John Blythe
Alan Gray
Francesca Willis
Dr. Emily Collins

SEPTEMBER 2020

Simulated Phishing and Employee Cybersecurity Behaviour (SPEC)

FULL REPORT

Dr. John Blythe, Alan Gray, Francesca Willis, & Dr. Emily Collins

This is the Full Report from the *Simulated Phishing and Employee Cybersecurity Behaviour (SPEC)* project, commissioned by CREST. The project investigate (i) how policies on simulated phishing emails are currently implemented in organisations using a cross-sectional survey and (ii) the impact of simulated phishing emails policies on employees' cyber security awareness and their perceptions of key factors (organisational trust, procedural fairness, stress and perceived monitoring) through an experimental study.

www.crestresearch.ac.uk/projects/spec

This research was funded by the Centre for Research and Evidence on Security Threats – an independent Centre commissioned by the Economic and Social Research Council (ESRC Award: ES/N009614/1) and which is funded in part by the UK Security and intelligence agencies and Home Office.

www.crestresearch.ac.uk

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The risk of cyber-attacks to UK companies is bigger than ever. With 90% of cyber breaches involving phishing techniques, it is increasingly important for organisations to identify ways to increase awareness of phishing attacks (NCSC, 2016) whilst maintaining positive relationships with employees (Kirlappos & Sasse, 2015). Not only is it important for organisations to know who is susceptible to what kinds of phishing attacks, but they also need to prevent such incidents from occurring. For this reason, a number of organisations conduct simulated phishing exercises, in which employees are sent emails that simulate phishing attempts. Organisations can use simulated phishing to test which employees are susceptible to what kinds of phishing attacks, provide instant feedback and timely, "just-in-time" training when links are clicked, and form the basis of repercussions for individuals who click phishing links.

Raising awareness might not be sufficient to successfully protect an organisation, especially if such exercises carry any unintended, negative outcomes. For instance, maintaining trust between employees and organisations is a vital component of compliance with security policies (Kirlappos & Sasse, 2015). Simulated phishing exercises have been argued to undermine this trust (Murdoch & Sasse, 2017), and create a hostile environment, whereby employees are blamed or actively punished for slip-ups, ultimately reducing long-term reporting (Murdoch & Sasse, 2017; NCSC, 2018). However, some of these assertions have not been directly tested, and do not account for the different ways in which simulating phishing could be implemented (e.g. to provide feedback, training or punishment to those falling victim).

In SPEC, we sought to address these research gaps through two studies. In study 1, we aimed to explore how organisations use simulated phishing and their use of "carrots" and "sticks" in their cyber security campaigns through a cross-sectional survey with awareness professionals. We found that organisations varied in their usage but found that sanctions are used in over 90% of organisations. We found that organisations took a stepped-response to repeat clickers in simulated phishing exercises, providing further intervention (such as further training or sanctions) depending upon susceptibility.

In study 2, we aimed to explore the impact of three different interventions: forced just-in-time (JIT) training (in which participants were forced to do a mandatory training course), lean JIT (a brief training message) and punishment (loss of performance payment) on anti-phishing detection and task performance. We also investigated the secondary impact on mental workload, perceived fairness, state anxiety and task autonomy. Through an experimental assessment, we found that all interventions significantly decreased phishing susceptibility. However, forced JIT negatively impacted task performance, fairness, state anxiety and mental workload. Punishment was also found to negatively impact fairness, state anxiety, and mental workload. These findings highlight that forced JIT and punishment should only be implemented in organisations with caution due to their negative consequences on employee wellbeing. Instead, the study encourages the use of lean JIT to decrease phishing susceptibility without negatively impacting mood, productivity or perceived fairness.

Overall, both studies highlight the need for greater consideration of behaviour change techniques (such as training and punishment) and their potential unintended consequences in organisations. By taking a "people-centric" approach to cyber security, one that sees people as a solution will be essential to building cyber resilience against phishing.

INTRODUCTION

Cyber security has been a tier 1 priority for the UK Government since 2011 (HM Government, 2016). Since then, cyber threats have massively increased: we have seen ransomware impact on our health services and data breaches become more and more common. Organisations, small and large, continue to struggle with managing cyber security risks.

The threat is very real. Cybercrime accounts for 50% of all crime in England and Wales (Office for National Statistics, 2018). With the introduction of GDPR, companies can now be fined up to twenty million pounds for inadequate data protection. Phishing continues to remain a prevalent security risk and is one of the most common ways in which access to company passwords and other sensitive information is gained. 88% of organisations in 2019 had been a victim of a phishing attack (Proofpoint, 2020) that routinely exploit the erroneous behaviour of employees (Blythe & Coventry, 2018). When successful, phishing attacks can result in reputational damage, monetary losses and the disruption of operational performance (e.g. Piggin, 2016; Zetter, 2016). Protecting against phishing attacks is, therefore, a key element in an organisation's cyber security initiatives.

Cyber criminals deliberately target employees with social engineering tactics to get users to click on fraudulent links, download malicious attachments or disclose sensitive information (Workman, Bommer, & Straub, 2008). It's unsurprising, then, that measures are being taken (and millions being spent) to manage human cyber risk. But some of these measures used by organisations are sparking concern - heavy monitoring and the use of metrics to target, and punish, 'at risk' employees (Caputo, Pfleeger, Freeman & Johnson, 2014; Kirlappos & Sasse, 2015; Sasse, 2015; Murdoch & Sasse, 2017; Reinfelder, Landwirth & Benenson, 2019). Indeed, more companies in the UK are beginning to adopt a culture of blame when it comes to their employees and cyber security (Ashenden & Sasse, 2013). Some security professionals, for instance, commonly endorse a belief in the employee as the "weakest link". They tend to place the responsibility of a breach upon individual users, rather than the organisation's culture, the quality of available training, or the design of security policies and procedures themselves (Adams & Sasse, 1999; Inglesant & Sasse, 2010; Parkin, van Moorsel, Inglesant, & Sasse, 2010; Zimmermann & Renaud, 2019).

When it comes to managing human cyber risk, security-awareness professionals rely on metrics. Metrics serve an essential function for decision makers in organisations: helping to assess risk over time and to aid strategic investment in cyber security resources--particularly if they focus on targeting the three key pillars of human cyber risk: Awareness, Behaviour and Culture (Coventry, Briggs, Blythe, & Tran, 2014). Most commonly, organisations gather metrics on training completion and click-rates on simulated phishing tests. This involves sending simulated emails to employees in an organisation and monitoring the 'click-rate' (i.e. how many people click on the link within the email) which may also be supplemented with "just-in-time" training. This approach gives employees training exactly when they fall for a phish.

Simulated phishing can serve three key purposes. Firstly, they provide organisations with metrics around how susceptible their workforce is to phishing attacks. CPNI (2017) proposes that simulated phishing can help organisations to understand whether some groups or departments are more susceptible to phishing emails, whether employees are susceptible to certain types of phishing threats and also whether susceptibility changes over time.

Secondly, simulated phishing can provide just-in-time feedback to employees if they fall for the phish with the potential to drive awareness and behaviour change.

Thirdly, they can also be a useful metric to evaluate the effectiveness of awareness training, offer additional support to vulnerable groups of people and help plan organisational cyber defences.

Simulated phishing can lead to behaviour change via its just-in-time training mechanism (Kumaraguru et al., 2017; Kumaraguru, Sheng, Acquisti, Cranor & Hong, 2010; Siadati, Palka, Siegel & McCoy, 2017), however, its effectiveness may depend on whether it is "lean" (i.e. short and does not interrupt productivity) or "rich" (i.e. requires doing a training course). It is the use of metrics derived from simulated phishing that raises concerns over its potential unintended consequences.

There has been a trend towards using such metrics to exclude, constrain and control staff (Zimmermann & Renaud, 2019). Some organisations use metrics gathered from simulated phishing to identify 'weak links' and punish them into online security: docking pay and locking computers until awareness programs have been completed and the user in question has been remedied and strengthened (Murdoch & Sasse, 2017; NCSC, 2018).

There have been questions regarding the unintended consequences of using simulated phishing harming the trust relationship between the company and employee and may reduce cyber security behaviour (Murdoch & Sasse, 2017) and other factors (such as trust) that are important for employee satisfaction and wellbeing (NCSC, 2018).

The NCSC (2018) found that there was a widespread blame culture around phishing with companies believing that if users were blamed for clicking on links, they would spot them next time. They also advise against using phishing simulations as a tool to catch people out as they may discourage reporting phishing emails in the future.

The implementation of Awareness, Behaviour and Culture activities including phishing simulation is often set by CISOS or other decision makers such as awareness professionals. It is recognised that CISOs often neglect the human aspect of cyber security (Ashenden & Sasse, 2013), creating restrictive policies that impede cyber security and organisational performance (Adams & Sasse, 1999; Inglesant & Sasse, 2010). Research has begun to understand how decision makers develop policies with human aspects and the cognitive biases that may lead to ineffective policies and how decision support systems can be used to debias potential decision making (Parkin et al., 2010).

However, there is no existing research on decision makers and simulated phishing emails and whether there is consideration of the potential impacts of different policy implementations (such as the role of rewards and punishments) on the human element. We seek to explore this within the SPEC project.

Despite discussion of the negative outcomes of simulated phishing, the majority of the work has been hypothetical. They have not been tested in research nor do they account for the many ways in which simulated phishing emails may be implemented in the wild and the types of policies (both positive and negative) that may be used. It is known that organisations differ in the way they implement and use simulated phishing emails.

These may range from simply warning a user that they clicked and explaining how they could have spotted the fake phish, to locking someone's IT until they complete a training course (NCSC, 2018), and reprimanding and disciplining employees who click (Murdoch & Sasse, 2017). These different implementations are likely to have positive effects on employee awareness but depending on the consequence of clicking may have the potential to impact negatively on employees (such as procedural fairness), wellbeing factors (such as stress) and productivity.

Although there has been an in-depth discussion of the utility of simulated phishing, there has been less

empirical research exploring the positive and negative aspects of different implementations. To address these research gaps, the SPEC project aimed to address two key overarching research questions:

1. What policies do organisations adopt on simulated phishing emails and why?
2. What impact do these different simulated phishing email policies have on positive and negative outcomes?

The SPEC project aimed to investigate: (i) how policies on simulated phishing emails are currently implemented in organisations and (ii) to explore the impact of simulated phishing email policies on employees' cyber security behaviour, productivity and psychological outcomes.

This was achieved through two studies:

Study 1:
A cross-sectional survey with awareness professionals to explore how simulated phishing campaigns are currently implemented in organisations.

Study 2:
An experimental study to assess the effects of different implementations of simulated phishing on employees' cyber security behaviour, productivity and their mood, procedural fairness, workload and autonomy.

Next, we cover the related work exploring simulated phishing and the role of rewards and punishment in organisations.

RELATED WORK

PUNISHMENT FOR BEHAVIOUR CHANGE

Punishment is an increasingly prevalent means by which organisations attempt to facilitate knowledge transfer. Simulated phishing provides them with an opportunity to test their employees and to easily identify "weak links". However, organisations may vary in how they deal with repeat clickers with many of them using sanctions. By this view, known as the Rational Choice Model (RCM), people are rational entities who commit a crime or wrongdoing only if the perceived benefits outweigh the costs (McCarthy, 2002; Becker, 1968). In order to prevent bad behaviour, according to RCM, it should be as simple as appealing to reason, for instance, by making the costs (e.g. punishment) outweigh the benefits (Bankston & Cramer, 1974). Employees, so the reasoning goes, will compare gains and losses when adhering to (or violating) security regulations (Bulgurcu, Cavusoglu, & Benbasat, 2010; Han, Kim, & Kim, 2017; Li et al., 2018).

Training, in this view, must recalibrate the perceived consequences of disobedience and reduce the cost of compliance (D'Arcy, Hovav, & Galletta, 2009). The educational element may thus be understood as increasing an employee's knowledge of the threat (and of its probability) whilst simultaneously demonstrating techniques that make it easier to be secure (thereby reducing effort-costs). Punishment is, of course, an integral part of this, or at least as far as the RCM assumes, and may be implemented alongside training to reinforce an employee's understanding. This can occur before or after the occurrence of a security incident and may have general or specific deterrence as its aim–seeking to reduce the likelihood of future security violations or discourage past violators from re-committing an offence, respectively (Siegel & Senna, 2008; Stafford & Warr, 1993).

Furthermore, it is thought that employees may be influenced by a sanction even if they are never personally its recipient. Simply observing or hearing of another's punishment may suffice to motivate compliance (cf. Bandura, 1977; Stafford & Warr, 1993). Afterall, the mere fear of penalties has been shown to decrease intentions to disobey information security regulations (Akers, Sellers, & Jennings, 2016; D'Arcy, Hovav, & Galletta, 2009; Son, 2011; Moody, Siponen, & Pahnila, 2018). Interestingly, observing the punishment of another (i.e. vicarious punishment) has been shown to influence future perceptions of actual sanctions, as they are experienced (Aurigemma & Mattson, 2017) – suggesting that the two in combination may evoke the greatest effects.

However, intentions are not the same as actions, and the actual efficacy of punishment as a means of increasing cyber security compliance remains unclear. Results are mixed, with studies showing performance improvements, detriments, and no effect at all. However, these studies rely on self-report measures exploring deterrence as means to predict compliance rather than an assessment of its effects (Chen et al., 2018; Cram, Proudfoot, & D'Arcy, 2017; Herath & Rao, 2009; Hovav & D'Arcy, 2012; Kuo et al., 2017; Peace, Galletta, & Thong, 2003; Siponen, Pahnila, & Mahmood, 2010).

A further explanation in line with deterrence theory (Gibbs, 1975; Pratt, Cullen, Blevins, Daigle, & Madensen, 2006), is that not all deterrents are created equal. Punishment requires various conditions in order to be effective, and these centre around how the punishment is perceived. Among the constructs presumed to be essential, are (1) celerity, (2) severity, and (3) certainty. Celerity refers to the swiftness of the punishment (D'Arcy & Herath, 2011); certainty: the perceived likelihood that wrongdoing will be met with punishment (Chen et al., 2012); and severity: the extent

to which the punishment is perceived to cause harm (Paternoster, 2010).

Kim, Lee and Kim (2019) made the distinction between general and specific deterrence. They argue that training acts as a general deterrent within organisations as employees gain knowledge and skills to reduce the likelihood of future security violations. From this perspective, misuse of information systems has been shown to be deterred through employees normative beliefs and self-efficacy (Pahnila & Mahmood 2010), motivation for punishment avoidance (Workman & Gathegi, 2007) and their perception sanction severity and certainty (D'Arcy, Hovav, & Galletta, 2009; Peace, Galletta, & Thong, 2003; Siponen, Pahnila, & Mahmood, 2010). The general deterrent effect of training is posited to occur through fear of penalties in the organisation. Specific deterrence, on the other hand, is often employed through the organisations stipulated punishment policies through sanctioning employees that violate policy. In line with deterrence theory, these are most effective if the sanction is swift, severe and certain.

There are very few studies that have explored the role of specific deterrents in cyber security. Earlier research by Harris and Furnell (2012) focussed on the role of shaming as punishment and found that shaming could have a positive influence but there are potential risks involved. Latest research by Kim, Lee and Kim (2019) explored the deterrent effects of both punishment and training on cyber security behaviour. Punishment was operationalised as a visit from the security team, loss of intranet access and threatened with a potential negative review in their performance appraisal for those employees who fell for simulated phishing. They found that the general deterrent of "training" and specific deterrent "punishment" were effective in reducing phishing susceptibility.

The opposite of punishment is, of course, rewards which is considered a more positive strategy. Rewards include financial, non-financial and psychological benefits provided to employees in return for their efforts (Bratton & Gold, 2017) and can be extrinsic (such as pay and benefits) or intrinsic (such as job fulfilment). Research has shown that reward strategies serve as a reliable mechanism for improving behaviour (Ajmal, Bashir, Abrar, Khan & Saqib, 2015), workplace trust (Burke, 2017) and work engagement (Jacobs, Renard, & Snelgar, 2014), even when the rewards are non-material (e.g. through gamification). However, the use of rewards employed within an organisational context for changing security behaviour has also received little investigation.

JUST-IN-TIME (JIT) TRAINING

Just-in-Time (JIT) training is the educational and training mechanism of simulated phishing. It is considered advantageous over traditional training as it helps trainers to retain and transfer knowledge more effectively (Al-Daeef et al., 2017). Unlike conventional training, JIT training is provided at the time the user needs it - when they fall for a phishing attack (Wash & Cooper, 2018) and is directly integrated into the primary tasks that employees perform at work (Kumaraguru et al., 2007).

As computers are used for a multitude of purposes in the workplace, security behaviours are considered secondary to primary work tasks. Dual-task interference posits that performing two simple tasks can interfere with each other if they are performed concurrently (Pashler, 1994). This phenomenon has been explained by the Capacity Sharing Model (Tombu & Jolicœur, 2003), which proposes that when two tasks (such as anti-phishing detection and work tasks) are performed simultaneously, performance on one task will be impaired as users have limited processing capacity. As a consequence, users are more likely to devote their cognitive processing to responding to work emails and may subsequently fall victim to phishing attempts.

Just-in-time training can overcome interference by presenting the training at critical times and through reducing the cognitive demand from users remembering how to be secure (Chueh & Barnett, 1997; Murtaugh et al., 2005). Instead, just-in-time training directs users to use sufficient processing capacity on secondary security tasks.

Additionally, research on context-dependent memory has proposed that embedded training is beneficial as it provides users with the opportunity to learn how to be secure in the context in which they would usually be attacked.

Specifically, psychologists have shown that users will be more likely to remember how to act when they are in the same context where they were trained (Anderson & Simon, 1996). Alternatively, it has been proposed that it is the immediate feedback from embedded training which increases the efficiency of learning (Schmidt & Bjork, 1992) and that users retain and transfer a greater amount of information from embedded training (Kumaraguru et al., 2007). T

he immediacy of feedback is also important for enhancing users' levels of response efficacy (i.e. by checking for phishing heuristics they can reduce phishing attempts). Response efficacy is a well-documented barrier to engagement in cyber security behaviour (Blythe, Coventry, & Little, 2015; Blythe & Coventry, 2018). JIT training is one use case of simulated phishing, the second is the use of metrics to identify vulnerable users for additional intervention.

PSYCHOLOGICAL IMPACT OF JIT TRAINING AND PUNISHMENT

Murdoch and Sasse (2017) argue that simulated phishing may harm the relationship between company and employee, and even ultimately reduce cybersecurity behaviour. Indeed, organisational trust has been shown to be an essential component for employee satisfaction, a key element of information-security policy compliance (Kirlappos & Sasse, 2015). And punishing employees for falling victim to simulated phishing emails is linked to frustration and resentment towards security staff (Caputo et al., 2014), both of which are likely to create problems when communicating future security messages (Ashenden & Sasse, 2013).

The UK's National Cyber Security Centre holds a similar view. They argue that simulated phishing is often used as means to blame and punish staff into behaving securely. The NCSC reasons that punishment for phishing is ineffective as individual differences and situational factors influence susceptibility and that sanctions do not change this. Further, they argue that sanctions can damage employees' trust, dissuade potential reporting and that it potentially reduces employee confidence in anti-phishing detection (NCSC, 2018).

Nevertheless, these claims of the possible long-term consequences of simulated phishing and punishment remain largely hypothetical and untested. Next, we cover potential psychological outcomes that may be impacted by different forms of simulated phishing.

Simulated phishing may result in unintended consequences on employees' productivity, as research has shown that email interruptions to primary tasks in the workplace negatively impact task performance (Addas & Pinsonneault, 2018; Monk, Trafton, & Boehm-Davis, 2008). As discussed, this can be explained by dual-task interference, which posits that performance on one or both tasks will be impaired when performed simultaneously due to humans' limited processing capacity (Pashler, 1994). Additionally, psychologists have proposed that the required effort to read lengthy just-in-time training uses significant processing capacity (Xiong, Proctor, Yang, & Li, 2019). Therefore, employees may have an increased mental workload (Moray, 1979). Research has found that when the demand of a task is greater than the resources available, performance will be significantly impaired

(Navon & Gopher, 1979). However, this relationship may be moderated by primary and interruption task variables, as research has found that complexity of the task and the dissimilarity of content between the primary and interruption task exacerbates the effect on performance (Edwards & Gronlund, 1998; Speier, Valacich, & Vessey, 1999). This effect may be the case with just in time training, where the content is dissimilar to employees' primary work tasks.

Nevertheless, just-in-time training may also reduce task autonomy. Task autonomy has been conceptualised as the state when an individual is given considerable independence, discretion and control over their task-at-hand (Hackman, 1980). Specifically, employees may experience decreased autonomy if they lack task-relevant experience in security (Chang, Huang & Choi). Research has shown that task autonomy has a positive motivational effect on task performance and productivity (Langfred & Moye, 2004: Joo, Jeung, & Yoon, 2010; Deci & Ryan, 2011). Therefore, it can be concluded that interruptions from phishing simulations which are irrelevant to primary work tasks, may negatively impact performance, workload perceptions and task autonomy in complex workplace tasks.

Moreover, existing work has focussed on the impact of task interruptions on performance on tasks such as serial recall tasks (Williams, Morgan, & Joinson, 2017; Morgan & Partick, 2013) which involve participants trying to remember a sequence of items (usually six to nine numbers, letters or both) and place high demand on cognitive processing. Brumby, Janssen and Mark (2019) note that interruption experiments are criticised for their lack of ecological validity as they often bear little resemblance to people's actual work environments and how they manage the interruptions that they experience in the workplace. We seek to overcome these limitations by utilising a platform that simulates work email-based tasks in a virtual email inbox.

Furthermore, phishing simulations may impact how people feel at the moment by increasing anxiety and stress levels amongst employees. When presented with threatening demands or dangers, people may experience unpleasant emotional arousal referred to as state anxiety. This psychological response, akin to feelings of nervousness, worry and stress bring about a sense of unease and stimulate the "fight or flight" adrenaline stress response (Kahn & Byosiere, 1992). Anxiety leads to worrying thoughts and attentional biases disrupting performance (Lukasik, Waris, Soveri, Lehtonen & Laine, 2019), with greater anxiety causing greater disruption.

The aforementioned value of punishment in behaviour change may be dependent on whether the sanctions are seen as 'just' and 'procedurally fair' within the workplace (Kassin, Fein, & Markus, 2014). Procedural fairness is an employee perception that a particular activity in which a person is involved in is conducted fairly (Lind, & Tyler, 1988). As described previously, there is already a concern regarding the fairness of punishment for cyber security behaviour at the strategic level (NCSC, 2018). Thus, we believe it's important to explore how the use of punishment within simulated phishing and JIT training itself may be viewed by users.

STUDY ONE

Human cyber risk management by security awareness professionals: Carrots or sticks to drive behaviour change?

INTRODUCTION

Metrics serve an important function for decision makers in organisations: helping to assess risk over time and to aid strategic investment in cyber security resources. However, there has been a trend towards using such metrics to exclude, constrain and control staff (Zimmermann & Renaud, 2019). Some organisations use metrics gathered from simulated phishing to identify 'weak links' and punish them into online security: docking pay and locking computers until awareness programs have been completed and the user in question has been remedied and strengthened (Murdoch & Sasse, 2017; NCSC, 2018).

Whilst simulated phishing does have benefits for enacting behaviour change through "just-in-time" training (Kumaraguru et al., 2017; Kumaraguru et al., 2010; Siadati et al., 2017), it is the use of metrics derived from simulated phishing that raises concerns over its potential unintended consequences.

A wealth of discussion has focused on the use of simulated phishing as means to punish staff with a focus on "blame culture", however, we do not know the various ways organisations use simulated phishing in their policies in practice. Moreover, we do not know how organisations use rewards and punishment more broadly. In study 1 we, therefore, aimed to explore how simulated phishing campaigns are currently implemented with a focus on how organisations deal with "repeat clickers". We also focussed on how rewards and sanctions are used to manage cyber security behaviour and what factors may influence such usage.

The use of rewards and sanctions in organisations will depend upon the personal and organisational resources available to security professionals (Ashenden & Sasse, 2013). While security professionals have traditionally realised their security aims by taking an authoritarian stance (Dhillon & Backhouse, 2001), the flattening of organisational structures has increased the need for security professionals to persuade, rather than coerce, employees toward secure behaviour (Ashenden & Sasse, 2013). To be successful here, within these new organisational structures, research in the field of management studies has suggested that an organisational "change agent" must possess expertise, credibility, political access to senior management and control of rewards and sanctions in order to be effective (Hardy, 1996).

These resources are necessary because any given change strategy can face resistance within the organisation from staff across all levels (Hardy, 1996). Unless an employee's experiences converge entirely with the strategist's, without adequate power (expertise, reward and sanction control), strategic initiatives will encounter disagreements and fail to be successfully executed by awareness professionals. Employees, after all, may agree on one end, but disagree on the means by which it may be achieved (Walsh, 1981). The awareness professionals' expertise and control of rewards and sanctions may therefore be key to managing human cyber risk but whether security professionals perceive themselves as possessing these resources and this sort of power, however, also remains unknown.

Viewing employees as part of the "problem" (Zimmermann & Renaud, 2019) in cyber security may also influence the degree of usage of rewards and

sanctions. Employees in cyber security are commonly referred to as the "weakest link" (Sasse, Brostoff, Weirich, 2001). As such, recent calls have focussed on viewing employees as part of the "solution" (Zimmermann & Renaud, 2019) - where we recognise that human errors do happen and that we not demonise employees but instead recognise that people are part of the system and their ability can have a positive role in cyber security. Whilst most research has looked at this mindset in security professionals (Ashenden & Sasse, 2013), there is little research on these influences on their use of rewards and sanctions.

AIMS AND RESEARCH QUESTIONS

Based on the existing work focusing on the impact of punishment on employee behaviour, particularly in the context of simulated phishing. The primary aim of study 1 was to explore how organisations use behavioural strategies of rewards and sanctions, as part of their cyber security awareness campaigns and how they deal with "repeat-clickers' as identified in phishing simulations. The second aim of this paper was to assess whether use of rewards and sanctions is influenced by security professionals' perceived control of rewards and sanctions, tendency to blame the user and their perceived impact of simulated phishing.

Using a cross-sectional online survey, the study 1 addressed the following research questions:

1. *What is the prevalence of reward and sanction use within organisations?*
2. *How do organisations approach the treatment of those who repeatedly click upon simulated phishing links?*
3. *What factors influence security professionals' use of rewards and sanctions?*

METHOD

PARTICIPANTS

Data was collected via Qualtrics between October 9th and November 29th, 2019. Security awareness professionals were recruited using social networking sites (LinkedIn, Twitter), and through CybSafe's customers, partners and contacts. Recruitment focused on those with responsibility for the '*Human Element of Cyber Security*' such as Information Security Awareness Managers, Cyber Security Education and Awareness Officers, and Cyber Security Education and Awareness Leads.

93 participants responded to the advert and clicked to participate in the study. 48 cases were excluded due to incomplete or missing data (of which 38 had completed the consent form but nothing else). The final dataset included 45 participants (19 male, 16 female, 1 prefer to self-describe, 2 prefer not to say, 7 missing; age range: under 25= 4, 25-34= 3 , 35-44= 11, 44-54= 13, over 55= 4, prefer not to say= 3, missing= 15), and was comprised of security awareness professionals working in either the public (12), private (23) or charity (1) sectors (prefer not to say= 2, missing= 7), from a range of organisation sizes (small <50 staff= 3, Medium between 50 and 249 staff= 4, large >250 staff= 28, prefer not to say= 3, missing= 7). 71% of participants used simulated phishing in their organisation.

To be included in the study, all participants were required to be aged 18 and have responsibility for security awareness within their organisation. To participate in the second part of the study on simulated phishing, all participants were required to use the tool in their organisation.

MEASURES

BEHAVIOUR CHANGE STRATEGIES

We developed a list of 11 potential strategies covering rewards and sanctions (see Appendix A), derived from security blogs advising on behaviour change and the Behaviour Change Technique Taxonomy (Michie et al., 2013). Respondents answered whether their organisation had used any of the methods within the last 12 months when it comes to managing human cyber risk and resilience. Answers were permitted via three mutually exclusive tick-boxes (Yes, No, I Don't Know).

Managerial incentives were defined as those practices employed to encourage a particular behaviour and encompassed both material and social rewards (e.g. gifts and public recognition). Sanctions, on the other hand, were aimed at discouraging an action, and again were understood in material and non-material terms (e.g. disciplinary warning, restriction of privileged access).

We chose not to refer to these as "incentives" and "sanctions" to reduce social desirability bias within responses. It is important to note that many sanctions within cyber security are not considered as such, but rather as further training devices, even if they may result in trouble with one's line manager and pay deductions.

With this in mind, sanctions were understood to be practices which sought to prevent a behaviour, either by direct punishment (material or otherwise) or by incurring mandatory effort beyond one's typical work role (e.g. locking computer until training is complete, enforced training resits etc.).

TREATMENT OF REPEAT CLICKERS

To assess how organisations dealt with "repeat-clickers" in simulated phishing exercises, we asked the following open-ended question: "*Your organisation discovers that an employee is repeatedly clicking on simulated phishing emails. What would your company currently do (if anything) and why*?"

REWARD AND COERCIVE POWER

Reward and coercive power were measured by subscales adapted from the Rahim Leader Power Inventory (RLPI) (Rahim, 1989). Participants were asked to consider the extent to which they agreed with a series of eleven statements on a seven-point scale, ranging from "strongly agree" to "strongly disagree". While the RLPI addressed perceived leader power from a subordinate employee's perspective (i.e. the perceived power of their line-manager), the present study measured a supervisor's own perceptions of power. Both reward and coercive power scales demonstrated high reliability: α=.852 and .879 respectively.

ATTITUDES TOWARDS USER AND PERCEIVED CONSEQUENCES OF SIMULATED PHISHING

The 'attitudes towards users' scale targeted the tendency to blame the users for clicking on a malicious link. This self-devised scale was composed of three items (e.g. "It is the responsibility of individual employees to avoid clicking on phishing links"). The scale showed low internal reliability: α=.622 (though c.f Nunnally, 1978, who suggests alpha values as low as .5 provide sufficient evidence of reliability).

The measure of perceived consequences of simulated phishing was a five-item scale, targeting the perceived consequences security professionals may associate with their organisation's use of simulated phishing (α=.912). The scale addresses some potential side effects that have been hypothesized to result from simulated phishing (Caputo, Pfleeger, Freeman, & Johnson, 2014; Kirlappos & Sasse, 2015; Murdoch & Sasse, 2017). These side effects include employee frustration, resentment towards security staff, loss of trust, and harm to employee morale. Example items include "Our simulated phishing policy is damaging to employee morale" and "My organisation's simulated

phishing policy harms the relationship between our company and its employees". See Appendix B for full scales.

PROCEDURE

Before beginning the study, full ethical approval was granted by CREST's and the School of Management at the University of Bath's ethics committee. All participants accessed the study via a link, recruited via email or social media platforms (e.g. LinkedIn, Facebook, Twitter). All those who clicked the link were first redirected to an information page, covering the study's rationale, what it required of participants, and how their data may be used in future, and consent was granted from participants wishing to take part. Following consent, participants were then asked a series of closed questions on the metrics their organisation recorded and the measures they took to encourage 'good' and reduce 'bad' cyber security behaviour. This was followed by questions on their perceived ability to control employee rewards and sanctions (i.e. coercive and reward power).

Participants were only taken to the second part of the study which focuses solely on phishing, and phishing simulations, if they indicated that their organisation uses simulated phishing metrics. They were first asked an open question regarding how they responded to repeat clickers. This was supplemented with a series of questions concerning the individual's perceptions of simulated phishing. The questionnaire ended with a series of demographic questions. Participants were debriefed on completion of the survey and informed of the overall aims of the project. They were reminded of their ability to withdraw their data at any time, and without need for explanation.

RESULTS

We conducted separate analyses for each research question. First, we explore the usage of behaviour change strategies and treatment of repeat clickers. We then explore which factors predict use of rewards and sanctions

Use of behaviour change strategies

The reporting frequency (%) of each behavioural change measure was assessed and the results of which can be seen in Figure 1.

Figure 1 shows that public recognition was the most common form of reward given to those demonstrating security behaviour (66% vs 63% certificates and 49% gift). The mean number of rewards was 4 (SD=2) and the mean number of punishments was 1 (SD=1). Overall, most respondents appeared to offer some form of reward: only 15% did not offer a single reward from the list, with 22% offering at least one of the listed rewards and 29.3% selecting all three. The most common form of punishment practiced from the list was informing an employee's line manager of risky behaviour (61%). The rarest, on the other hand, was naming and shaming an employee for risky behaviour (15%) and locking an employee's workstation until awareness training is complete (17%). Overall, most respondents appeared to offer some form of punishment: only 10% did not administer a single punishment from the list, with 20% administering at least one of the listed punishments and 24% selecting five or more.

TREATMENT OF "REPEAT-CLICKERS"

The open-ended responses were analysed using thematic analysis, following guidelines by Braun and Clarke (2006). Three key themes were drawn: '*Grading the response', 'Tailoring the training' and 'Non-punitive emphasis'.*

Grading the response: Many reported a stepped response to repeat clickers (27%), with clickers being

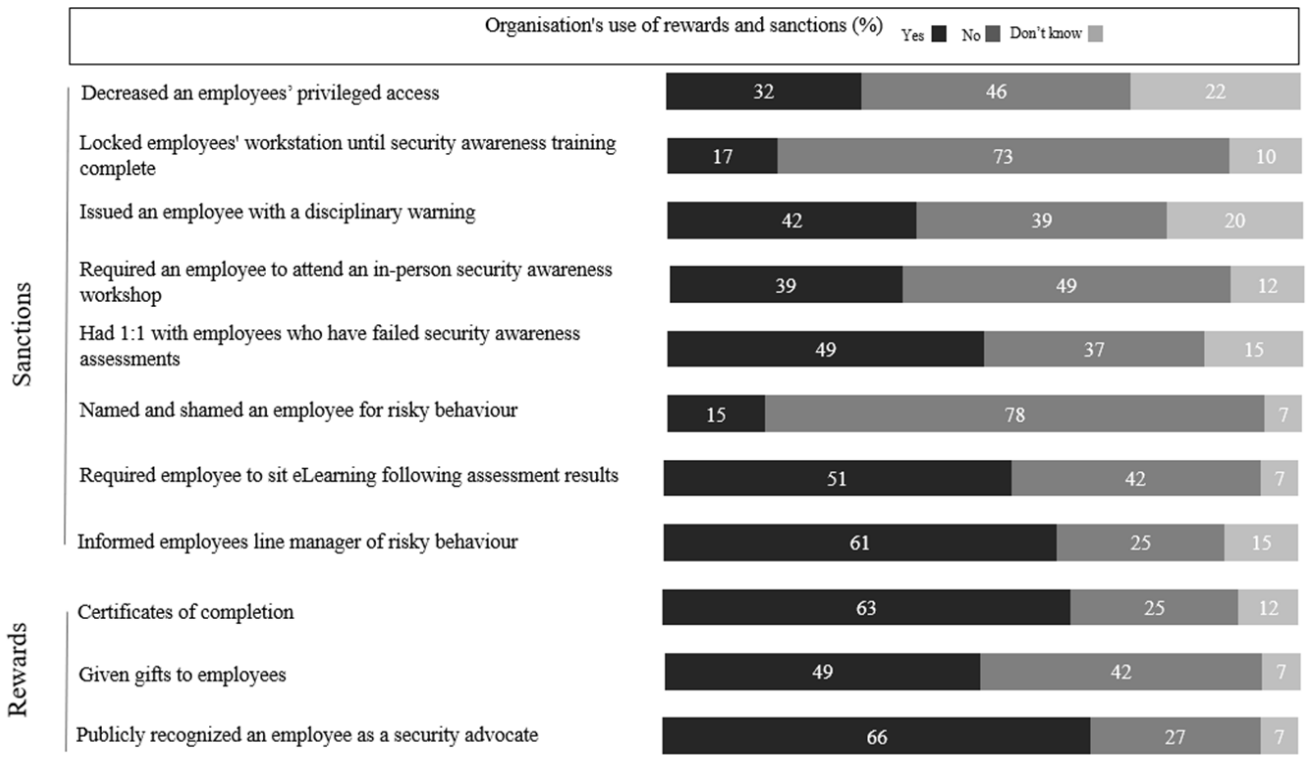


Figure 1. Reporting frequency (%) of rewards and sanctions

regularly re-phished and allocated further, more intensive training (or punishment) with each additional failing. For example,

"We assign additional awareness training as the first step. Then there is a sit-down meeting with a regional information security officer as the next step … limiting of a person's access to email, web, etc. as possible next consequences"

Training often begins via the learning management system (LMS) and progresses, with repeat offences, to 1:1 meetings and consultations (with security officers, managers, and CISOs). This is frequently accompanied, in the later stages, with an invitation to discuss the problem with their line-manager, where disciplinary action may be taken; one respondent even mentioned the possibility of discharge.

Tailoring the training: Many respondents mention various attempts to tailor training to the individual (36%). This seems to be achieved by either focusing the training topic around the type of phishing email or the role served by the employee within the company. Interestingly, within those who tailor training, there seems to be an attempt to understand clicking by looking to factors beyond the individual. Meetings and discussions are used to formulate a potential cause for the error (e.g. phishing type; role vulnerability) ---and these are used to design targeted treatments (e.g. role-specific training to detect where current awareness instruction is failing). For example,

"Each click comes with point in time education, repeat offenders must take re-training. We feel that repeat clicking is indicative of lack of knowledge needed to identify phishing emails, so we try to equip them to do better. If after meeting with a specific individual or group it seems like there is something about their specific role that makes them more susceptible to clicking, we try to work with them to design role specific training."

Non-punitive emphasis: Several respondents emphasised taking a "non-punitive" approach to remediating clickers (36%). Perhaps based on the view that "clicking is indicative of a lack of knowledge" —-and therefore potentially a difficulty at a 'macro', rather than individual level---respondents recognised the importance of being "nice about it". An "open culture" was understood as necessary for preventing 'people hiding or covering up issues', as well as for creating a context for "long-term change". For example,

"Talk to them as an education and training intervention, rather than a punitive one."

"Show them the error and how they were caught out. Show them the way to avoid it in the future and be nice about it. Everyone makes mistakes and an open culture is required to prevent people from hiding or covering issues us because they have made a mistake."

PREDICTING USE OF REWARDS AND SANCTIONS

DESCRIPTIVE STATISTICS

The means and standard deviations for the control of rewards and sanctions' measures and participants' perceptions of the perceived consequences of simulated phishing and tendency to blame users are presented in Table 1. In order to obtain a more easily interpreted percentage value, reward and punishment rates were calculated for each respondent according to the following two formulae:

Reward Rate = *(total number of rewards used ÷ total number of rewards listed) x 100)*

Punishment Rate = *(total number of punishing behaviours used ÷ total number of punishing behaviours listed) x 100).*

CORRELATIONS AMONG THE SCALES

Correlations between the five scales were then calculated. A significant positive correlation was found between reward power and coercive power (r=.784, $p<0.01$ [n=39]), suggesting that as perceived reward power increased so too did perceived coercive power. There was also a significant positive correlation between reward rate and punishment rate (r= .314, $p<0.05$ [n= 41]), suggesting that as reward rate increased so too did punishment rate. No further significant correlations were found among the five scales (e.g. tendency to blame the user, perceived consequences of simulated phishing etc.).

A multiple linear regression was then conducted to predict 'Punishment Rate' from 'Perceived Consequences of Simulated Phishing', 'Tendency to Blame the User', 'Reward Rate', 'Reward Power', and 'Coercive Power'. A non-significant regression equation was found ($F_{5,25}$=.335.898, $p>.005$, R^2=.092).

An additional multiple linear regression was then calculated to predict 'Reward Rate' from 'Perceived Consequences of Simulated Phishing', 'Tendency to Blame the User', 'Punishment Rate', 'Reward Power', and 'Coercive Power'. Again, this was non-significant ($F_{5,25}$=727.013, $p>.005$, R^2=.109.).

Variable	*Mean*	*SD*	*N=*
Reward power (1-7)	4.24	1.44	39
Coercive power (1-7)	3.31	1.54	39
Tendency to blame the user (1-7)	3.57	1.13	32
Perceived consequences of simulated phishing	2.99	1.45	32
Reward rate (1-100)	59.35	34.57	41
Punishment rate (1-100)	38.11	26.36	41

Table 1. Means and Standard deviations for total scores on leader power (reward and coercive), tendency to blame the user, perceived consequences of simulated phishing, and reward and punishment rates.

DISCUSSION

In study 1 we aimed to explore how organisations use rewards and sanctions as part of their cyber security awareness campaigns, and how they deal with "repeat-clickers' as identified in phishing simulations. We also assessed whether the use of rewards and sanctions is influenced by security professional's perceived control of rewards and sanctions, tendency to blame the user and their perceived impact of simulated phishing. We capture our interpretation of the findings in more detail in the next section, before moving on to a discussion of the work's limitations, implications, and conclusion.

BEHAVIOUR CHANGE STRATEGIES

Managerial incentives and sanctions were readily reported among our respondents, with all organisations using at least one of the strategies listed and many employing the full range. These findings demonstrate that organisations use a combination of strategies to deliver behaviour change, with variability in the types of sanctions employed by organisations. As acknowledged in the introduction, such strategies, at least as far as punishment is concerned, may prove counterproductive: decreasing an employee's morale, organisational trust, and even willingness to engage in cyber security behaviour (Caputo, Pfleeger, Freeman, & Johnson, 2014; Kirlappos & Sasse, 2015; Murdoch & Sasse, 2017).

While less concern has been expressed about rewards in cyber security management, there is also reason to believe that they too could pose a negative impact: replacing intrinsic motivation with its inferior extrinsic counterpart (Boss, Kirsch, Angermeier, Shingler, & Boss, 2009; Patterson, Grenny, Maxfield, McMillan, & Switzler, 2008) and removing security behaviours that aren't rewarded or monitored in some way. Furthermore, rewarding what should be 'normal security behaviour' might make employees see it as exceptionally good behaviour, and something that is not generally achieved. This said, however, claims regarding the efficacy of reward and punishment in cyber security remain largely hypothetical and untested (Patterson et al., 2008; Siponen, Willison & Baskerville, 2008; Warkentin & Willison, 2009).

Study 1 also uncovered substantial variation in the prevalence of individual reward and punishment strategies. Public recognition was the most common form of reward given to those demonstrating strong cyber security behaviour---with newsletter mentions and certificates being appreciably more likely than material gifts. While the efficacy of varying reward types remains to be tested (Cf. Harris & Furnell, 2012; Aurigemma & Mattson, 2017), it is likely that the value of public recognition in this context depends strongly on an organisation's cyber security culture, and the individual's internalisation of those values (Han, Kim & Kim, 2017).

Similarly, the most common form of punishment was informing an employee's line manager of risky behaviour. Again, this strategy will rely on the cyber security culture within which it is executed, and the line manager's understanding and respect for cyber security in general. It is hard, for instance, to imagine such reporting leading to much change if the line managers themselves are ignorant of the value of cyber security or consider it a frustration and mere bureaucracy.

Many who employ punishment techniques, to be clear, do not understand them as such: preferring to understand these methods within the rubric of 'training,' rather than sanctions per se (Siponen et al., 2008); and while all effort was made to couch questions with this in mind, several practices may have been dismissed because of injudicious lapses in language sensitivity. The rarest sanction, for example, was 'naming and shaming an employee for risky behaviour', with just 15% of respondents reporting its use.

TREATMENT OF REPEAT OFFENDERS

Considering the treatment of repeat clickers, re: simulated phishing, thematic analysis revealed three key themes: "*grading the response*", "*tailoring the training*", and "*non-punitive emphasis*".

Training was the most common recourse--- though what form that training took was varied (e.g. LMS, 1:1, educational videos) and often graded (in severity) for repeat clickers, with clear efforts to tailor the treatment and a frequent emphasis on non-punitive methods. This attempt at acknowledging the needs of the individual, rather than unrolling blanket policy when it comes to at-risk employees, demonstrates that organisations are tailoring their approaches. Many participants used a graded response to dealing with repeat clickers---with the severity of the sanction increasing as click rate increased.

Most organisations used simulated phishing as a means to deliver more tailored training to their staff. Simulated phishing allows organisations to identify "high-risk" employees, giving them more targeted training that may be specific to the role or individual needs.

Although organisations did employ a range of sanctions, 36% of participants emphasized a "non-punitive" approach to remediating clickers, with few participants acknowledging the need for an open culture around cyber security risks. These findings suggest that whilst organisations are employing sanctions, there is a trend towards security awareness professionals understanding the need for a more humanistic approach.

PREDICTING USE OF REWARDS AND SANCTIONS

Correlational analysis revealed two significant positive correlations: the first between perceived reward and perceived coercive power, the second between reported reward rate and reported punishment. This suggests that rewarding good cyber security behaviour within an organisation is often paired with punishing bad behaviour. In other words, control over carrots is often paired with control over sticks, both in terms of the awareness professional's perceived power to execute rewards and punishments, and in their use by the company as a whole.

The role of punishment in cyber security is routinely cautioned (Murdoch & Sasse, 2017; NCSC, 2018). It has been suggested that its value as a tool for behaviour change may fluctuate depending on whether or not the sanctions are seen as 'just' and 'procedurally fair' within the workplace (Kim et al., 2019). Punishment is counterproductive in security. It leads to under reporting, resentment towards IT staff and impacts on trust and productivity (Sasse, 2015). A common mistake with simulated phishing is that it is used to play "gotcha" with employees – acting as means to entrap employees into security (Krebs, 2019). The current study supports the assumption that organisations do use punishment measures as means to change behaviour. Previous research has been largely anecdotal (Sasse, 2015; Murdoch & Sasse, 2017), and the current study highlights the prevalence of punishment use in organisations.

Contrary to the Hardy model, resource factors hypothesized to be critical in the production of an effective change agent (Hardy, 1996) did not predict the use of behaviour change strategies. Reward and coercive powers failed to predict either reward or punishment rates, suggesting that control over these resources may not guarantee their use, and that mere resource disposal may be necessary but not sufficient in the creation of an effective change agent. On the other hand, the role of the "security awareness professional " encompasses different types of job roles and levels of seniority. As the cyber security sector continues to mature and become more professionalised (UK Government, 2019), this diversity is expected to continue. This may mean that there is variability in awareness's professionals' control of rewards and sanctions due to

their position in the company and level of seniority. For example, whilst awareness professionals themselves may not believe in the use of punishment, the use of such policies may be dictated by senior management and the board of directors (who themselves have their own mental models around human cyber risk (Hinna, De Nito, Mangia, Scarozza & Tomo, 2014). Research in management studies has shown that political access to senior management is important for "change agents" in organisations (Hardy, 1996). Future research should therefore look at this access to senior management and also explore variability in the security awareness professional role and how their control of rewards and sanctions may be constrained by organisational structures and senior management influence.

LIMITATIONS

It should be mentioned that the practice of punishing employees is sufficiently stigmatised to raise concerns over the reliability of the present data. While all participants were repeatedly informed of their anonymity, and no explicit mention of 'punishment' was made to reduce social desirability bias, professionals may have shied away from specifying the particulars of their sanctioning strategies for fear of litigation or indicting their organisations.

The present study limits its scope to the perceptions and practices of professionals and ignores the effect of these practices on the employees themselves. It would, however, be interesting to expand this study to include the perceptions of the on-the-ground workers themselves. While much literature has assumed that simulated phishing practices are detrimental to employee morale and organisational trust, there is currently no work addressing this assumption beyond various small case-studies (Baldwin, Ford & Blume, 2017).

Another key concern for future research is understanding how professionals decide on a particular security campaign. The present work addresses the conclusions of professionals, in terms of their recording and managerial strategies, but not how they arrived at those decisions---ignoring the rationale at play, and the pressures potentially imposed by senior management.

Finally, we looked at the prevalence of behaviour change strategies as they pertained to rewards and sanctions. For behaviour change strategies to be effective, however, they need to target the drivers and barriers to security behaviour (Michie, Van Stralen, & West, 2011). Different strategies are more or less effective depending on whether it is a lack of capability, motivation and/or opportunity preventing the security behaviour (Michie et al., 2011). For effective behaviour change, the choice of strategies should therefore be guided by evidence and behaviour change frameworks (Michie et al., 2011). However, a "behavioural science" informed approach will depend on the capability of the security awareness professionals. Future work should explore the extent to which organisations' behaviour change interventions are based on behavioural science.

SUMMARY AND CONCLUSIONS

Study 1 showed that organisations vary widely in the "carrot" and "sticks" deployed. It highlights the need for a greater consideration of the human element of cyber security and demonstrates that punishment is widely used to manage human cyber risk. The use of rewards and punishments to promote cyber security is practiced to bridge the 'knowing-doing gap' in cyber security awareness, wherein knowledge of best practice is seldom met with adherence and actual security. In study 2, we seek to assess the impact of punishment on behaviour change and its potential impact on productivity and unintended consequences.

Contrary to the 'user as the weakest link' view, users do wish to protect their organisations and are often hindered from doing so by unusable security measures and policies (Sasse, 2015). This forms what research has described as a 'cycle of bad security' (Reinfelder et al., 2019); starting with a negative view of the user

that is ultimately vindicated by excluding them from security policy design and by failing to consider their everyday experience.

Security awareness professionals play an important role in managing human cyber risk. We found diversity in professionals' views and use of "punishment" as a tool but changing awareness professionals' beliefs about the human element and utility of behaviour change strategies will be key in addressing human cyber risk (Zimmermann & Renaud, 2019). We found that some professionals do recognise the need for a "non-punitive" approach but further guidance and recognition of this within industry standards and frameworks should be considered.

STUDY TWO

An experimental assessment of Just-in-Time training and punishment on productivity, anti-phishing detection and psychological outcomes

INTRODUCTION

Study 1 showed that organisations varied in their use of rewards and sanctions. We found that organisations differed in how they dealt with at-risk employees (i.e. repeat-clickers in simulated phishing) and used a graded approach to interventions that ranged from additional training to punishment. Existing work by Murdoch & Sasse (2017) and recommendations by the NCSC (2018) have argued that organisations use simulated phishing to blame and punish employees into behaving securely, but this may lead to unintended consequences. However, an experimental assessment is needed to explore this impact on positive and negative outcomes.

Building on from study 1 and existing research, we aimed to explore the impact of three different types of simulated phishing interventions. We designed the interventions to mimic the impact of simulated phishing in the workplace: lean JIT (a brief training message), forced JIT (whose task is locked until they complete a mandatory training course), and punishment (who are informed they lose a bonus payment).

We aimed to explore the impact of the interventions on participant's primary task performance and anti-phishing detection (susceptibility and report rate) compared to a control group. The extent to which JIT training and punishment impacts on psychological outcomes of mood, workload, fairness and autonomy is also investigated. Associated hypotheses are as follows:

H1: There will be a significant difference in task performance for those exposed to just in time training and punishment conditions compared to the control condition.

H2: There will be a significant difference in anti-phishing detection for those exposed to just in time training and punishment conditions compared to the control condition.

H3: Individuals exposed to just in time training and punishment conditions will elicit a significantly stronger response on psychological outcomes (mood, workload, fairness and autonomy) than those in the control condition.

METHOD

PARTICIPANTS

The sample consisted of 206 adults (71 males and 135 females), ranging between 18 and 74 years of years of age (M_{age} = 34, SD_{age} = 12). Participants were recruited from an online recruitment platform (prolific.ac). To be eligible for the study, participants were from the UK, had access to Chrome as a browser and consented to take part in studies that involved deception. Participants were paid £3.50 for their participation and an ostensibly "bonus" payment of £1.00, which they were told was dependent upon task performance, but was actually used as part of the manipulation and awarded to all).

Participants demographics by condition were: control (51 in total: 14 males, 37 females, M_{age} = 34, SD_{age} = 13), lean just-in-time (53 in total: 15 males, 38 females, M_{age} = 35, SD_{age} = 11), forced just-in-time (51 in total: 19 males, 32 females, M_{age} = 33, SD_{age} = 11) and punishment (51 in total: 23 males, 28 females, M_{age} = 34, SD_{age} = 12).

DESIGN

A between subjects design was used. Participants took part in a simulated office environment, completing office-based tasks in a virtual inbox and were subjected to both genuine task emails and simulated phishing emails. Upon clicking on a simulated phishing email, participants were randomly exposed to one of four conditions varying in the level of just-in-time training and punishment received (control, lean JIT, forced JIT, and punishment). The dependent variables were task performance, anti-phishing detection and psychological outcomes (mood, workload, fairness and autonomy).

MATERIALS

PHISHTRAY SETTINGS

The office environment was mimicked using the PHISHTRAY platform – an open source e-tray software (Joinson, 2019), that allows for people to be exposed to security threats in a sandboxed but immersive work-like experience. The platform provides greater ecological validity and overcomes a key limitation of existing phishing studies referred to as "subject expectancy effect" (in which people are more suspicious and biased towards looking for phishing emails as they have been prompted to do so by the experimenter).

PHISHTRAY allows the manipulation of the cognitive pressure that the user experiences (e.g., the time available, the number of emails, the presence of distractions, interruptions or competing goals) mimicking the productivity pressures that people experience in the real world.

Participants take on the role of a procurement manager called Geoff who has to deal with a series of tasks within a time limit (e.g. selecting a new supplier, dealing with complaints, dealing with a team dispute). Participants work through a series of emails where they have to make decisions based on the available information (see appendix C for example).

For the current study, we set the time limit for the exercise to 15 minutes with 21 task emails and 8 phishing emails. Task emails covered a range of tasks from checking suppliers' notes, information about team members and paying invoices. The simulated phishing emails were designed based on phishing heuristics and were varied in difficulty (e.g. presence of spelling mistakes, incorrect sender address, addressed directly to Geoff). The content of the phishing emails included invoices, password resets, file sharing emails and social media and were designed based on existing phishing research (e.g. Kleitman, Law, & Kay, 2018) (see appendix D for example).

Next, we cover how the conditions were implemented within the phishtray paradigm. With the exception of the control group, the manipulation appeared on phishing emails following 4 mins 30 seconds. Responses to genuine emails and phishing emails prior to this acted as a baseline measure.

In the lean JIT condition, if participants clicked on a phishing link they were directed to a single webpage on phishtray that informed them that they clicked on a phishing link and outlined the risks of phishing and three steps they can take to protect themselves (1. Were you expecting it, 2. Check the sender details and 3. Check the destination) (see Figure 4).

In the forced JIT, if participants clicked on a phishing link they were directed to a single webpage that informed them that their task was locked until they completed a mandatory training course. They were able to unlock the interface with a code following training completion (see Figure 2.) The course consisted of five pages which educated participants on social engineering and phishing emails. This was followed by four questions on the content of the training.

In the punishment condition, after they clicked on a phishing link, participants were directed to a single webpage that informed them that they had received a loss of payment (see Figure 3). Participants in the control condition were not presented with training or

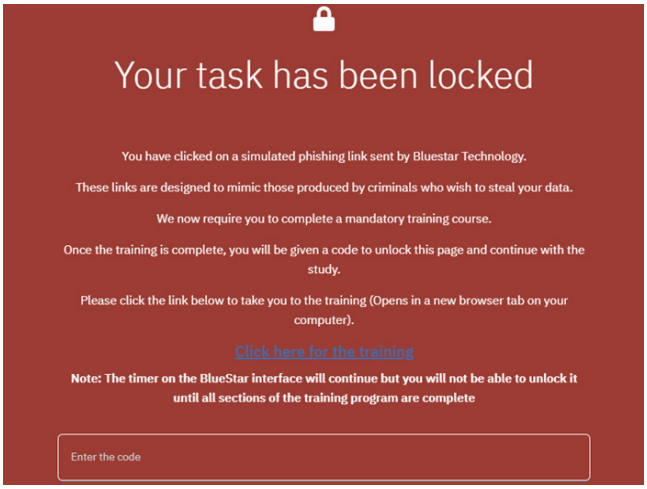


Figure 2. Forced JIT condition

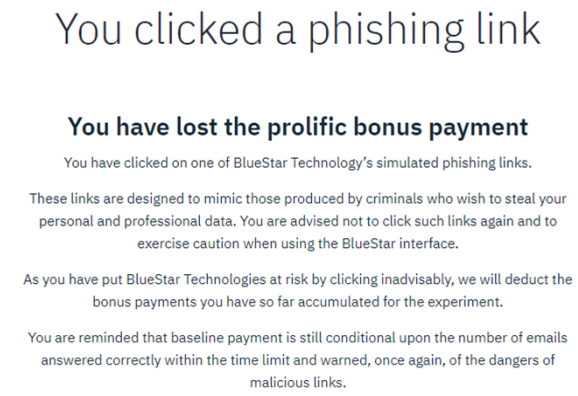


Figure 3. Punishment condition

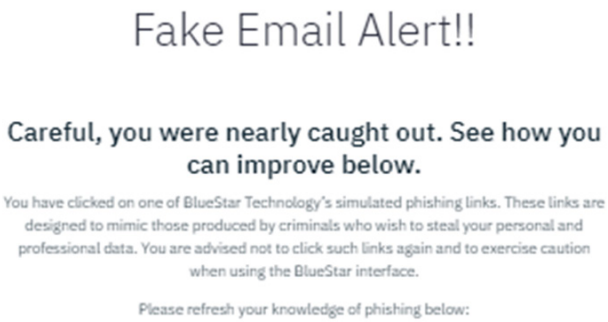


Figure 4. Lean JIT condition

a punishment message and were informed that they clicked on a phishing link.

For any email (both genuine and phishing), participants' responses were recorded as one of the following: no response, clicked a link or attachment, clicked a response option related to the task, reported the email or deleted the email.

PHISHTRAY BEHAVIOURAL MEASURES

The following measures were recorded from the phishtray platform:

- Task accuracy (the number of correct responses to genuine emails in the e-tray task)
- Phishing susceptibility (the number of phishing emails that participants responded to)
- Report-rate (the number of phishing emails participants reported)

PSYCHOLOGICAL MEASURES

MOOD

Mood was assessed using the 6-item short form of the state anxiety scale (STAI: Marteau & Bekker, 1992) in which participants report their subjective feelings of worry, tension and apprehension. An additional two items were added for individual assessment of stress and happiness ('I feel happy'; 'I feel stressed') (Hare, 2016). Participants recorded their current feelings on a 4-point Likert scale from 'not at all' to 'very much' with higher scores representing worse mood. The internal reliability of scale was $\alpha = .87$.

WORKLOAD

The 6-item NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) was used to assess workload across the dimensions of mental demand, physical demand, temporal demand, performance, effort and frustration. Participants rate how they feel across the 6 dimensions (e.g. '*How mentally demanding was the task*?') from '*very low*' to '*very high*' with the exception of performance ('perfect' to 'failure'). The response scale was VAS scale with 21 points displayed as a line to participants with higher scores representing greater workload. The internal reliability of scale was $\alpha = .79$.

PROCEDURAL FAIRNESS

The 5-item procedural fairness scale was adapted from Lind, Kanfer & Earley (1990) which covered participants' perceptions of procedural fairness (e.g. '*I found the task's processes fair*', '*I found that task fair*'). Participants rated each statement on a 5-point Likert scale from '*strongly disagree*' to '*strongly agree*'. The internal reliability of scale was $\alpha = .82$.

AUTONOMY

Task autonomy was measured on a 3-item scale adapted from Crane (2012) in which the outcome was changed to fit the current task (e.g. '*I felt that I had a choice in how to manage the emails and workload*'). Participants rated each statement on a 5-point Likert scale from '*strongly disagree*' to '*strongly agree*'. The internal reliability of scale was $\alpha = .76$.

PROCEDURE

Full ethical approval was granted by CREST's and the School of Management at the University of Bath's ethics committee.

The study was hosted online at Qualtrics.com and emtray.com which was accessible via the recruitment platform (prolific.ac). Participants were instructed to complete the survey via a desktop or laptop computer using the Chrome browser to allow the phishtray platform to display correctly across the screen.

Participants satisfying the eligibility criteria were invited to take part in an "Email Management" study exploring how people communicate when under pressure from ongoing tasks. All those who clicked the link were first redirected to an information page, covering the study's rationale, what it required of participants, and how their data may be used in future, and consent was granted from participants wishing to take part.

Participants were then required to watch a 3-minute video that explained how to use the task interface (participants were not able to progress past this page until they had watched the video).

Participants were then instructed that they would receive £3.50 for participation in the study but may also receive a bonus payment of an extra £1 for exceptional performance: measured by the number of tasks accurately completed within the 15-minute time limit. They were instructed to make sure they were to be quick and accurate as possible. Prior to starting the "emtray" interface, they were required to complete the baseline mood questionnaire.

Upon being directed to the emtray website, they were

provided with more information about the fictional organisation (Bluestar Technologies), their job role as Procurement Manager and required to read an Acceptable Usage and security policy - included to mimic organisation security policies around phishing emails and to watch out for scams and phishing attacks. They were then reminded to be as quick and accurate as possible.

Participants then had 15 minutes to answer as many emails as quickly and accurately as possible.

Upon completion of the emtray task, participants were then redirected to Qualtrics to complete questionnaires assessing mood, workload, fairness and autonomy. Participants were then debriefed about the study with detailed explanation of the deception used.

All participants received the bonus payment (that was "withdrawn" in the punishment condition) regardless of actual task performance.

We included three attention checks to ensure that participants were adequately paying attention throughout the study. The first and third were presented during the mood questionnaire at baseline and post-task in which participants had to select "somewhat agree" for one of the statements. For the second attention check, we assessed the number of correct responses within the emtray task.

RESULTS

TREATMENT OF DATA

The data from phishtray and that from Qualtrics were collated for analysis. Repeated measures ANOVA were conducted to assess the effect of simulated phishing conditions on task performance and anti-phishing detection. One-way ANOVAs were conducted to assess the effect of simulated phishing conditions on psychological outcomes, each of which will be discussed separately below. To account for differences in baseline mood, an ANCOVA was also conducted to explore the impact on mood. Issues with data storage on the phishtray platform meant that 23 participants' data was corrupted, however, we still retain these participants for assessing the impact on psychological measures

BEHAVIOURAL MEASURES

TASK PERFORMANCE

To determine the effects of the intervention on task performance, participants' overall percentage for task accuracy was taken at time 1 and time 2. Data was analysed using a mixed 4 (condition: control, lean JIT, forced JIT and punishment) x 2 (Time: T1/T2) ANOVA, with the condition as the between-subjects factor and time as the within subject factor. The analysis revealed a significant main effect of time on ability $F_{1,176}$=45.259, p<.001, with a large effect size partial η^2 = .205 and a significant interaction between time and condition $F_{3,176}$=3.373, p=.020, with a small effect size partial η^2 = .054. Bonferroni-corrected follow-up analyses revealed that there was a significant reduction in performance from T1 to T2 for all conditions (p<001). For comparisons between conditions, there were no significant differences between conditions at baseline (p=.583) but there were significant differences between conditions at time 2. Performance for individuals in the Forced JIT condition was significantly lower at time 2 compared to control (p=.028), lean JIT (p=.016) and punishment (p=.046). However, there were no other significant differences in performance at T2 for conditions.

PHISHING SUSCEPTIBILITY

Variable	Time	Control	Lean JIT	Forced JIT	Punishment
Task Accuracy	T1	93.38% (13.49%)	94.02% (10.56%)	92.10% (11.62%)	91.18% (13.54%)
	T2	86.79% (12.34%)	87.51% (13.07%)	76.93% (21.12%)	86.34% (13.84%)
Phishing Susceptibility	T1	94.48% (14.26%)	94.36% (14.39%)	84.51% (26.26%)	90.93% (24.22%)
	T2	81.67% (27.31%)	47.66% (32.18%)	12.56% (13.11%)	38.64% (31.52%)
Report rate		4.61% (14.92%)	16.72% (26.52)	11.50% (23.57%)	22.08% (32.40%)

Table 2. Means (SD) for behavioural indices

To determine the effects of the intervention on phishing susceptibility, participants' overall percentage for task accuracy was taken at time 1 and time 2. Data was analysed using a mixed 4 (condition: control, lean JIT, forced JIT and punishment) x 2 (Time: T1/T2) ANOVA, with the condition as the between-subjects factor and time as the within subject factor.

The analysis revealed a significant main effect of time on ability $F_{1,178}$=414.007, p<.001, with a large effect size partial η^2 = .699 and a significant interaction between time and condition $F_{3,178}$=30.041, p<.001, with a large effect size partial η^2 = .336.

Bonferroni-corrected follow-up analyses revealed that there were no significant differences between conditions at baseline (p=.062) but there were significant differences between conditions at time 2. Forced JIT was significantly lower at time 2 compared to all other conditions (p<.001), punishment and lean JIT were both significantly lower than the control group (p<.001) but there was no significant difference between them (p=.345). There was a significant reduction in performance from T1 to T2 for all conditions (p<001).

REPORT RATE

There was a statistically significant difference between conditions on report rate as determined by a one-way ANOVA $F_{3,181}$=4.107, p<.01, partial η^2=.064. As equal variances could not be assumed, a Games-Howell post hoc test revealed that report rate was significantly higher for punishment (M=22.08%, p<.01) and lean JIT (M=16.72%, p=.034) compared to control (M = 4.61%). There was no difference in reporting rate between forced JIT (M=11.50%) and control, and there were no other statistically significant differences between conditions.

PSYCHOLOGICAL MEASURES

MOOD

There was a statistically significant difference between conditions on perceived mood as determined by a one-way ANCOVA whilst controlling for baseline mood $F_{3,201}$=8.734, p<.001, partial η^2=.115.Bonferroni-corrected post hoc tests showed there was a significant difference between control and forced JIT (p=.005) and punishment (p<.001) but not for the lean JIT (p=1.000). The significant difference was also consistent between lean JIT and forced JIT (p=.044) and punishment (p=.002). There was no significant difference in mood between locked and punishment (p=1.000).

Comparing the estimated marginal means showed that mood was worse for punishment (M=46.46), followed by locked (M=44.75), lean JIT (M= 40.05) and control (M=38.84).

WORKLOAD

There was a statistically significant difference between conditions on perceived workload as determined by

Variable	Control	Lean JIT	Forced JIT	Punishment
Mood (0-100)	38.84 (8.84)	40.05 (8.85)	44.75 (8.83)	46.46 (8.86)
Workload (0-100)	30.42 (16.46)	37.77 (19.24)	46.14 (20.31)	41.73 (17.78)
Fairness (1-5)	4.21 (.53)	3.92 (.61)	3.65 (.86)	3.53 (.77)
Autonomy (1-5)	3.75 (.83)	3.52 (.84)	3.43 (.88)	3.44 (.82)

Table 3. Mean (SD) for psychological indices

a one-way ANOVA $F_{3,\ 202}$=6.629, p< 0.001, partial η^2=0.090. A Tukey post hoc test revealed that perceived workload was significantly higher for the forced JIT condition (M=46.14, p<.001) and punishment condition (M=41.73, p=.012) compared to the control group (M=30.42) (but not for lean JIT; M=37.77, p=.183). There were no other statistically significant differences between conditions.

FAIRNESS

There was a statistically significant difference between conditions on perceived fairness as determined by a one-way ANOVA $F_{3,202}$= 9.651, p<.001, partial η^2=.125. As equal variances could not be assumed, a Games-Howell post hoc test revealed that perceived fairness was significantly lower for punishment (M=3.53, p<.001) and forced JIT (M= 3.65, p<.01) when compared to control (M=4.21). Perceived fairness was not significantly lower for lean JIT (M= 3.92, p=.056) when compared to control (M= 4.21). Perceived fairness for punishment was also significantly lower compared to lean JIT (p=.021) but not for forced JIT. There were no other statistically significant differences between conditions.

AUTONOMY

There was no statistically significant difference between conditions on perceived autonomy as determined by a one-way ANOVA $F_{3,202}$=1.518, p=.211, partial η^2=.022.

DISCUSSION

Using a novel methodology, study 2 explored the impact of different types of simulated phishing interventions compared to a control group. We manipulated the type of training administered and the presence of punishment. First, we explored the impact on their task performance and anti-phishing detection ability. Secondly, we explored the impact on psychological outcomes of mood, workload, fairness and autonomy.

Our first hypothesis was that there would be a significant reduction in task performance for those exposed to training and punishment compared to control (H1). We found that there was only a significant impact of forced JIT on task accuracy. Those individuals who were forced to do a training course that prevented them from doing their primary tasks, had worse task accuracy. This supports existing work that emails interruptions to primary tasks impacts negatively on task performance (Addas & Pinsonneault, 2018; Monk, Trafton, & Boehm-Davis, 2008). In line with dual-task interference, forced lengthy JIT training uses significant processing capacity that it impacts primary task performance (Xiong, Proctor, Yang, & Li, 2019; Pashler, 1994). Compared to lean JIT, forced training is more complex and may have greater dissimilarity of content between the primary and interruption task which is known to exacerbate the effect on performance (Edwards & Gronlund, 1998; Speier, Valacich, & Vessey, 1999). Lean JIT, on the other hand, which is short and may not require significant processing capacity, is better linked to primary tasks and therefore, not impact on primary task performance.

The second hypothesis was that there would be a significant difference in anti-phishing detection for those exposed to JIT training and punishment conditions compared to the control condition. We found that forced JIT was most effective, having a significantly reduced phishing susceptibility compared to other interventions. Lean JIT and punishment were equally effective in reducing susceptibility compared to the control group. There was also a difference in report rate, with those exposed to punishment or lean JIT reporting significantly more phishing emails compared to control. However, this effect was not observed for forced JIT. These findings support the dual task theory that just-in-time training can directs users to use sufficient processing capacity on anti-phishing detection (Tombu & Jolicœur, 2003), increases the efficiency of learning (Schmidt & Bjork, 1992) and may help users retain and transfer a greater amount of information from embedded training (Kumaraguru et al., 2007). Forced JIT did not lead to increased reporting, but lean JIT and punishment did. As forced JIT impacted task accuracy, it may have led people to engage in more risk-averse behaviour - choosing not to engage with proactive cyber security behaviour (such as reporting). Existing work has argued that fear of penalties and punishment may impact reporting behaviour (NCSC, 2018). However, we note that this effect was not observed for the punishment condition.

From a general deterrence perspective (Gibbs 1975; Pratt, Cullen, Blevins, Daigle, & Madensen, 2006), both JIT training and punishment act as deterrents to phishing. It has been argued that JIT training acts as a general deterrent, helping users to gain knowledge and skills to detect phishing attacks. On the other hand, specific deterrents like punishment, are effective if swift, severe and certain. We found that punishment did not add any additional benefits beyond the general deterrent effect of training. This supports Kim et al. (2019) who found both JIT training and punishment to be effective in reducing phishing susceptibility, however, their study does not draw comparisons between their relative effectiveness. The current study showed that punishment was not as effective as forced JIT.

The third hypothesis was that individuals exposed to JIT training and punishment conditions will elicit a significantly stronger response on psychological

outcomes than those in the control condition. Whilst forced JIT had a significantly more effective impact on anti-phishing detection, we found several impacts on psychological outcomes.

Firstly, we found that participants in the forced JIT and punishment condition had significantly worse mood than those in lean JIT and the control group. These findings suggest that being forced to do training and being punished for clicking on simulated phishing links can significantly increase people's state anxiety. State anxiety is a form of psychological stress and is an important factor in employee well-being. When people experience state anxiety, it can lead to long-term impacts on employee satisfaction (Zalewska, 2011). When people experience negative emotions associated with cyber security, it undermines people's confidence in cyber security (NCSC, 2018) and impacts on intrinsic trust between employees and organisations (Kirlappos & Sasse, 2015). This unintended outcome therefore has the potential to negatively impact security and compliance more broadly.

Secondly, we found that participants in the forced JIT and punishment condition had significantly higher mental workload than the control group (but not the lean JIT). Mental workload can be subjectively felt by the individual who perceives a cost while realising a task (Hart & Staveland, 1988), with research showing that when the demand of a task is greater than the resources available, performance will be significantly impaired (Navon & Gopher, 1979). The findings suggest that punishing clickers or forcing users to do training increases mental demand on users. This increased mental demand also impacts on task accuracy, particularly when the intervention significantly disrupts an individual's primary task as in the forced JIT condition. Long term exposure to increased mental demand increases the likelihood of errors in many different professions (e.g. Yurko, Scerbo, Prabhu, Acker, & Stefanidis, 2010) and can contribute to employee burnout and stress (e.g. Xiaoming, Ma, Chang, & Shieh, 2014).

Thirdly, perceived fairness was significantly lower for the forced JIT and punishment condition compared to the control. Perceived fairness for punishment was also significantly lower compared to lean JIT. According to theories of punishment, whilst a sanction must be seen as swift, severe and certain to be effective it must also be seen as "just" and "fair" by people (Kassin, Fein, & Markus, 2014). We found that locking tasks to force training and punishing participants was not seen as procedurally fair by users.

Combined, these findings support the assertions by Murdoch and Sasse (2017) and the NCSC (2018) that simulated phishing can have negative effects. However, we find that there are differences in the level of impact depending upon how simulated phishing is administered. For lean JIT, when a gentle reminder is given to participants about what to look out for, the JIT training is effective on anti-phishing detection but does not lead to negative impacts on productivity, mood and workload. Furthermore, lean JIT is seen as equally procedurally fair as a control group. Whilst locking users' tasks for training or punishing clickers does reduce susceptibility, it negatively impacts on workload and mood and is not viewed as procedurally fair by participants.

This study is one of the first to experimentally explore the potential unintended consequences of simulated phishing. Future research would benefit from exploring the impact within an organisational setting. Factors such as reduced productivity, mood and increased workload are likely to lead to long-term impacts on employee satisfaction and morale (Zalewska, 2011; Xiaoming, Ma, Chang, & Shieh, 2014). Moreover, future studies could further look at the impact on the stress response. We looked at psychological stress and found that forced JIT and punishment increased participants' state anxiety. However, exploring the physiological response to stress requires measuring heart rate and cortisol reactivity (Paxion, Galy, & Berthelon, 2014). Further work could look at whether JIT training and punishment leads to the activation of the body's physiological response to

stress.

Further research could also explore the role of other factors that influence susceptibility. Individual differences such as personality and risk preference are known to influence susceptibility to phishing (Williams, Beardmore & Joinson, 2017) as is the influence mechanism that is used within the phishing email. Existing work has found that nudges can help to attenuate the impact of traits such as risk-taking and impulsivity in the context of cookie acceptance (Coventry, Jeske, Blythe, Turland, & Briggs, 2016). Further work could look to account for individuals' differences to explore what role they play in the effectiveness of simulated phishing and the impact of traits can be reduced through JIT training.

LIMITATIONS

Of course, this study was not without its limitations. Firstly, as we varied the level of task interruptions to reflect real-world use of simulated phishing, we were not able to explore the impact of these interruptions on resumption lag (Brumby, Janssen, & Mark, 2019). Resumption lag is often used in interruptions research to explore the time taken to recover from an interruption and the extent to which this leads to errors. Resumption lag is taken to reflect the time that is needlessly "wasted" as a consequence of being interrupted. We were interested in whether people accurately responded to emails, rather than their performance on cognitively demanding working-memory based tasks. However, future research should look to explore the impact of JIT training and punishment on resumption lag. Secondly, we did not look at the long-term impact of the interventions so we cannot make an assessment on long-term behaviour change as used within Kim, Lee, and Kim (2019).

SUMMARY AND CONCLUSIONS

In this study, we have shown that different types of just-in-time training and use of punishment from phishing simulations can have different impacts. Specifically, we found that two forms of JIT training (lean and forced) and punishment significantly improved anti-phishing detection comparative to control. Forced JIT with more in-depth training is the most effective in reducing phishing susceptibility. However, despite these benefits, we found that there were several negative impacts of forced JIT training and punishment on people's mood, workload and perceived fairness.

We conclude that one form of JIT (lean JIT) potentially provides the most benefits to users as it was effective on anti-phishing detection and did not impact on productivity, mood, workload and perceived fairness. The policy implications of this seem, at this stage, fairly clear. Decisions to use simulated phishing should be guided by their use as an educational and behaviour change tool that provides immediate feedback (but must remain short and not overly disruptive on user's primary work tasks). When used as a means to force users to do training or provide punishment is unlikely to yield any additional benefits to reducing risk but may cause unintended consequences to psychological well-being. Organisations should therefore be cautious in how they use simulated phishing as a mechanism, focussing more on its utility for providing JIT feedback and less of an opportunity to play "gotcha" with employees.

SPEC CONCLUSIONS

Phishing attacks are a significant threat to organisational cyber security that exploit employee behaviour. Managing the uncertainty of phishing attacks brings a significant challenge to organisations. Simulated phishing is used by organisations to train staff and also to identify individuals that may require additional intervention such as training or punishment. We explored what policies organisations adopt on simulated phishing emails and the potential impact on positive and negative outcomes.

We found that organisations vary widely in the "carrot" and "sticks" deployed. It highlights the need for a greater consideration of the human element of cyber security and demonstrates that punishment is widely used to manage human cyber risk. The use of rewards and punishments to promote cyber security is practiced to bridge the 'knowing-doing gap' in cyber security awareness, wherein knowledge of best practice is seldom met with adherence and actual security. We found that many adopted a stepped response to repeat clickers, with clickers being regularly re-phished and allocated further, more intensive training (or punishment) with each additional failing.

Security awareness professionals play an important role in managing human cyber risk. We found diversity in professionals' views and use of "punishment" as a tool but changing awareness professionals' beliefs about the human element and utility of behaviour change strategies will be key in addressing human cyber risk (Zimmermann & Renaud, 2019). We found that some professionals do recognise the need for a "non-punitive" approach but further guidance and recognition of this within industry standards and frameworks should be considered.

In our experimental study, we found that forcing training on users and using punishment negatively impacts on psychological well-being. We found that simulated phishing does offer training and behaviour change benefits but can have secondary consequences depending on how it is enacted. It highlights the need for organisations to consider how they use such practices to manage human behaviour. We argue that there needs to be a step-change towards developing "*Just and Fair*" cultures in organisations (NCSC, 2018) which focus on security accountability between leaders and staff and drop notions of blame. Researchers refer to this as a paradigm shift in cyber security towards focusing on "humans as a solution" (Zimmermann & Renaud, 2019). By acknowledging the complexity and interconnectedness of cyber security within the workplace and viewing humans as contributors will lead to greater cyber resilience in organisations.

REFERENCES

Adams, A. & Sasse, A. (1999). Users are not the enemy. *Communications of the ACM, 42*(12), 40-46.

Addas, S. & Pinsonneault, A. (2018). E-mail interruptions and individual performance: Is there a silver lining? *MIS Quarterly, 42,* 381-406.

Ajmal, A., Bashir, M., Abrar, M., Khan, M., & Saqib, S. (2015). The effects of intrinsic and extrinsic rewards on employee attitudes; mediating role of perceived organisational support. *Journal of Service Science and Management, 8*(4), 461.

Akers, R. L., Sellers, C. S., & Jennings, W. G. (2016). *Deterrence and rational choice theories. Criminological Theories: Introduction, Evaluation, and Application,* 7th ed, 14–43. New York, NY: Oxford University Press.

Al-Daeff, M. M., Basir, N., & Saudi, M. M. (2017). Security Awareness Training: A review. *Proceedings of the World Congress on Engineering 2017, 1,* 1-6.

Anderson, J. R., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25,* 5-11.

Ashenden, D., & Sasse, A. (2013). CISCOs and organisational culture: Their own worst enemy? *Computers and Security, 39,* 396-405.

Aurigemma, S., & Mattson, T. (2017). Deterrence and Punishment Experience Impacts on ISP Compliance Attitudes. *Information and Computer Security, 25*(4), 421–436.

Baldwin, T. T., Ford, J. K., & Blume, B. D. (2017). The state of transfer of training research: Moving toward more consumer-centric inquiry. *Human Resource Development Quarterly, 28*(1), 17–28.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bankston, W. B., & Cramer, J. A. (1974). Toward a macro-sociological interpretation of general deterrence. *Criminology, 12,* 251–280.

Becker, G. S. (1968). Crime and punishment: An economic approach. *The economic dimensions of crime, Journal of Political Economy 76*(2), 169-217.

Blythe, J. M., & Coventry, L. (2018). Costly but effective: Comparing the factors that influence employee anti-malware behaviours. *Computers in Human Behavior*, *87*, 87–97. https://doi.org/10.1016/j.chb.2018.05.023.

Boss, S. R., Kirsch, L. J., Angermeier, I., Shingler, R. A., & Boss, R. W. (2009). If someone is watching, I'll do what I'm asked: Mandatoriness, control, and information security. *European Journal of Information Systems, 18,* 151-164.

Bratton, J., & Gold, J. (2017). *Human resource management: Theory and practice.* Palgrave.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Brumby, D. P., Janssen, C. P., & Mark, G. (2019). How do interruptions affect productivity? In C. Sadowski & T. Zimmermann (Eds.), *Rethinking Productivity in Software Engineering* (pp. 85-107). Apress, Berkeley, CA.

Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly, 34(*3), 523-548.

Burke, W. W. (2017). *organisation change: Theory and practice.* Sage publications.

Caputo, D. D, Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2014). Going spear phishing: Exploring embedded training and awareness. *IEEE Security and Privacy, 12*(1), 28-38.

Chang, J. W., Huang, D. W., & Choi, J. N. (2012). Is task autonomy beneficial for creativity? Prior task experience and self-control as boundary conditions. *Social Behavior and Personality, 40,* 705-724.

Chen, Y., Ramamurthy, K., & Wen, K. W. (2012). Organizations' information security policy compliance: Stick or carrot approach? *Journal of Management Information Systems, 29(3),* 157-188.

Chen, X., Wu, D., Chen, L., & Teng, J. K. (2018). Sanction severity and employees' information security policy compliance: Investigating mediating, moderating, and control variables. *Information & Management, 55(8)*, 1049-1060.

Chueh, H., & Barnett, G. (1997). "Just-in-Time" Clinical Information. *Journal of the Association of American Medical Colleges, 72,* 512.

Coventry, L., Briggs, P., Blythe, J., & Tran, M. (2014). *Using behavioural insights to improve the public's use of cyber security practices.* London: Government Office for Science.

Cram, W.A., Proudfoot, J., & D'Arcy, J. (2017). Seeing the forest and the trees: A meta-analysis of information security policy compliance literature. *Innovative Behavioral IS Security and Privacy Research*, 4051-4060.

CPNI. (2017). *Phishing Simulations Guide*. Available at https://www.cpni.gov.uk/system/files/documents/51/d7/phishing_simulations_guide.pdf

D'Arcy, J., & Herath, T. (2011). A review and analysis of deterrence theory in the IS security literature: Making sense of the disparate findings. *European Journal of Information Systems, 20*(6), 643-658.

D'Arcy, J., Hovav, A., & Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: A deterrence approach. *Information Systems Research, 20*(1), 79-98.

Deci, E. L., & Ryan, R. M. (2011). Self-Determination Theory. *Handbook of Theories of Social Psychology, 1,* 416-433.

Dhillon, G., & Backhouse, J. (2001). Current directions in IS security research: towards sociotechnical perspectives. *Information Systems Journal, Blackwell, 11*(2), 127-153.

Edwards, M. B., & Gronlund, S. D. (1998). Task interruption and its effects on memory. *Memory, 6,* 665-687.

Gibbs, J. P. (1975). Crime, punishment, and deterrence. *The Southwestern Social Science Quarterly, 48*(4), 515-530.

Hackman, J. R. (1980). Work redesign and motivation. *Professional Psychology: Research and Practice, 11,* 445– 455.

Han, J., Kim, Y. J., & Kim, H. (2017). An integrative model of information security policy compliance with psychological contract: Examining a bilateral perspective. *Computers & Security, 66,* 52–65.

Hardy, C. (1996). Understanding power: 'Bringing about strategic change'. *British Journal of Management (Special issue), 17,* S3-S16.

Harris, M., & Furnell, S. (2012). Routes to security compliance: Be good or be shamed? *Computer Fraud and Security, 12,* 12–20.

Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organisations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems, 47*(2), 154-165.

Hinna, A., De Nito, E., Mangia, G., Scarozza, D., & Tomo, A. (2014). Advancing public governance research: Individual and collective dynamics in and around the boardroom. *Studies in Public and Non-Profit Governance, 2*, 3–39.

HM Government. (2016). *National Cyber Security Strategy 2016-2021*. London: HM Government.

Hovav, A., & D'Arcy, J. (2012). Applying an extended model of deterrence across cultures: An investigation of information systems misuse in the US and South Korea. *Information & Management, 49*(2) 99-110.

Inglesant, P., & Sasse, M. (2010). The true cost of unusable password policies: Password use in the wild. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell., R. Jeffries, & G. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 383-392).

Jacobs, S., Renard, M., & Snelgar, R. J. (2014). Intrinsic rewards and work engagement in the South African work retail industry. *SA Journal of Industrial Psychology, 40*(2), 1-13.

Joo, B., Jeung, C., & Yoon, H. J. (2010). Investigating the influences of core self evaluations, job autonomy, and intrinsic motivation on in-role job performance. *Human Resource Development Quarterly, 21,* 353-370.

Kim, B., Lee, D., & Kim, B. (2019). Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. *Behaviour & Information Technology,* 1-20.

Kirlappos, I., & Sasse, M. A. (2015). Fixing security together: Leveraging trust relationships to improve security in organisations. *Proceedings of the NDSS Symposium 2015* (1), 1-10.

Kahn, R. L., & Byosiere, P. (1992). Stress in organisations. In M. D. Dunnette & L. M Hugh (eds.), *Handbook of Industrial and organisational Psychology* (pp.571-650). Palo Alto, CA:

Krebs, B. (2019). *Should failing phishing tests be a fireable offense?* Retrieved April 30, 2020, from https://krebsonsecurity.com/2019/05/should-failing-phish-tests-be-a-fireable-offense.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2017). Protecting people from phishing: The design and evaluation of an embedded training email system. *Proceedings of the ACM CHI 2007 Conference on Human Factors in Computing Systems, 1,* 905-914.

Kumaraguru, L., Rhee, Y., Sheng, S. Hasan, S., Acquisti, A., Cranor, L., & Hong, J. (2007). Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. e-Crime *Researchers Summit, Anti-Phishing Working Group.* Doi: 10.1145/1299015.1299022.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology, 10*(2), 1-31.

Kuo, K., Talley, P. C., Hung, M., & Chen, Y. (2017). A deterrence approach to regulate nurses' compliance with electronic medical records privacy policy. Systems-level quality improvement, *Journal of Medical Systems, 41*(12), 1-10.

Langfred, C. W., & Moye, N. A. (2004). Effects of Task Autonomy on Performance: An Extended Model Considering Motivational, Informational, and Structural Mechanisms. *Journal of Applied Psychology, 89,* 934-945.

Lind, E, A., & Tyler, T.R. (1988). *The Social Psychology of Procedural Justice.* New York: Plenum Press.

Li, H., Luo, X. R., Zhang, J., & Sarathy, R. (2018). Self-control, organizational context, and rational choice in Internet abuses at work. *Information & Management, 55(3)*, 358-367.

Lukasik, K. M., Waris, O., Soveri, A., Lehtonen, M., & Laine, M. (2019). The relationship of anxiety and stress with working memory performance in a large non-depressed sample. *Frontiers in Psychology, 10,* 4.

McCarthy, B. (2002). New economics of sociological criminology. *Annual Review of Sociology, 28,* 417-442.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine, 46*(1), 81-95.

Michie, S., Van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation science, 6*(1), 42.

Monk, C. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied, 14*(4), 299-313,

Moody, G. D., Siponen, M., & Pahnila, S. (2018). Toward a unified model of information security policy compliance. *MIS Quarterly, 42*(1), 285-311.

Moray, N. (1979). *Mental workload: Its theory and measurement*. New York: Plenum.

Morgan, P. L., & Patrick, J. (2013). Paying the price works: increasing goal access cost improves problem solving and mitigates the effect of interruption. *The Quarterly Journal of Experimental Psychology, 66*(1), 160-178,

Murdoch, S. J., & Sasse, M. A. (2017). Should you really phish your own employees? Retrieved 30 April, 2020, from https://tech.newstatesman.com/business/phishing-employees.

Murtaugh, C. M., Pezzin, L. E., McDonald, M. V., Feldman, P. H., & Peng, T. R. (2005). Just-in-Time Evidence-Based E-Mail "Reminders" in Home Health Care: Impact on Nurse Practices. *Health Services Research, 40,* 849-864.

Navon, D., & Gopher, D. (1979). On the economy of the human processing systems. *Psychological Review, 86,* 254–255.

NCSC. (2018). The trouble with phishing. Retrieved 30 April, 2020, from tps://www.ncsc.gov.uk/blog-post/trouble-phishing.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Office for National Statistics. (2018). *Crime in England and Wales: year ending March 2018*. London: ONS.

Parkin, S., van Moorsel, A., Inglesant, P., & Sasse, M. (2010). A stealth approach to usable security: Helping IT security managers to identify workable security solutions. In A. Keromytis, & S. Peisert (Eds.), Proceedings of the 2010 workshop on new security paradigms (pp. 33-49).

Pashler, H. (1994). Dual-Task Interference in Simple Tasks: Data and Theory. *Psychological Bulletin, 116,* 220-244.

Paternoster, R. (2010). How much do we really know about criminal deterrence? *Journal of Criminal Law and Criminology, 100*(3), 765-824.

Patterson, K., Grenny, J., Maxfield, D., McMillan, R., & Switzler, A. (2008). *Influencer: The power to change anything.* New York, NY: McGraw-Hill.

Peace, A. G., Galletta, D. F., & Thong, J. Y. L. (2003). Software piracy in the workplace: A model and empirical test. *Journal of Management Information Systems, 20*(1), 153-177.

Piggin, R. (2016). Cyber security trends: What should keep CEOs awake. *International Journal of Critical Infrastructure Protection*, *13*, 36–38. Retrieved from https://www.infona.pl/resource/bwmeta1.element.elsevier-11cbc911-9de2-3e8a-9927-cbb49c221304.

Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madesen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. In F. T. Cullen (Eds.), *Taking Stock: The Status of Criminological Theory.* (pp.367-395). Routledge.

Proofpoint. (2020). State of the phish report. Retrieved from https://www.proofpoint.com/uk/resources/threat-reports/state-of-phish.

Rahim, A. M. (1989). Relationships of leader power to compliance and satisfaction with supervision: Evidence from a national sample of managers. *Journal of management, 12*(4), 545-556.

Reinfelder, L., Landwirth, R., & Benenson, Z. (2019). Security managers are not the enemy either. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, p.433. ACM (2019).

Sasse, A. (2015). Scaring and bullying people into security won't work. *IEEE Security & Privacy, 13*(3), 80-83.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal, 19*(3), 122-131.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217.

Siadati, H., Palka, S., Siegel, A., & McCoy, D. (2017). *Measuring the effectiveness of embedded phishing exercises.* In: 10th USENIX Workshop on Cyber Security Experimentation and Test.

Siegel, L. J., & Senna, J.J. (2008). *Introduction to Criminal Justice*, 11th ed. Belmont, CA: Wadsworth.

Siponen, M., Pahnila, S., & Mahmood, A. M. (2010). Compliance with information security policies: An empirical investigation. *Computer, 43*(2), 64-71.

Siponen, M, Willison, R., & Baskerville, R. (2008). Power and Practice in Information Systems Security Research." *Proceedings of the International Conference on Information Systems,* 1–12. Paris, France: Association for Information Systems.

Son, J. (2011). Out of fear or desire? Toward a better understanding of employees' motivation to follow IS security policies. *Information & Management, 48*(7), 296-302.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences, 30,* 337-360.

Stafford, M. C., & Warr, M. (1993). A reconceptualization of general and specific deterrence. *Journal of Research in Crime and Delinquency, 30,* 123-135.

Tombu, M., & Jolicœur, P. (2003). A Central Capacity Sharing Model of Dual-Task Performance. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 3-18.

UK Government. (2019). *Developing the UK cyber security profession.* Retrieved from https://www.gov.uk/government/consultations/developing-the-uk-cyber-security-profession.

Wash, R., & Cooper, M. M. (2018). Who provides phishing training? Facts, stories, and people like me. *CHI 2018 Honourable Mention, 492*, 1-12.

Walsh, C. (1981). Power and advantage in organisations. *Organisation Studies, 2*(2), 131-152.

Warkentin, M., & Willison, R. (2009). Behavioral and Policy Issues in Information Systems Security: *The Insider Threat. European Journal of Information Systems, 18*(2), 101–105.

Williams, E. J., Morgan, P. L., & Joinson, A. N. (2017). Press accept to update now: Individual differences in susceptibility to malevolent interruptions. *Decision Support Systems, 96,* 119-129.

Workman, M., Bommer, W., & Straub, D. (2008). Security lapses and the omission of information security measures: A threat control model and empirical test. *Computers in Human Behavior, 24*(6), 2799–2816. Doi: 10.1016/j.chb.2008.04.005.

Workman, M., & Gathegi, J. (2007). Punishment and ethics deterrents: A study of insider security contravention. *Journal of the American Society for Information Science and Technology, 58*(2), 212-222.

Xiong, A., Proctor, R. W., Yang, W., & Li, N. (2019). Embedding training within warnings improves skills of identifying phishing webpages. *Human Factors and Ergonomics Society, 61,* 577- 595.

Zetter, K. (2016). *Inside the cunning, unprecedented hack of Ukraine's power grid.* Retrieved from https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/

Zimmermann, V., & Renaud, K. (2019). Moving from a "human-as-problem" to a "human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies, 131,* 169-187.

APPENDICES

APPENDIX A

BEHAVIOUR CHANGE STRATEGIES SCALE

To the best of your knowledge, has your organisation used any of the following within the last 12 months when it comes to managing human cyber risk and resilience? *(Yes/No/I Don't know)*

- Publicly recognised an employee as a security advocate (e.g. in an organisational newsletter, email etc.)
- Given gifts to employees (e.g. prize draw, vouchers, time off)
- Informed an employee's line manager of risky behaviour (e.g. non-course completion, failing a phishing test)
- Certificates of completion (e.g. awareness course completion
- Required an employee to sit/resit e-learning following assessment results
- Named and shamed an employee for risky behaviour
- Had a 1:1 with employees who have failed security awareness assessments
- Required an employee to attend an in-person security awareness workshop
- Issued an employee with a disciplinary warning
- Locked an employee's work station until security awareness training is complete
- Decreased an employee's privileged access
- Other (please specify): _____

APPENDIX B

ATTITUDES TOWARDS USERS' SCALE

Please indicate the extent to which you agree with the following statements (Strongly Agree - Strongly Disagree)

- It is the responsibility of individual employees to avoid clicking on phishing links
- Employees who click on simulated phishing links should be punished
- It is wrong to blame employees who click on simulated phishing links

Perceived consequences of simulated phishing

Please indicate the extent to which you agree with the following statements (Strongly Agree - Strongly Disagree)

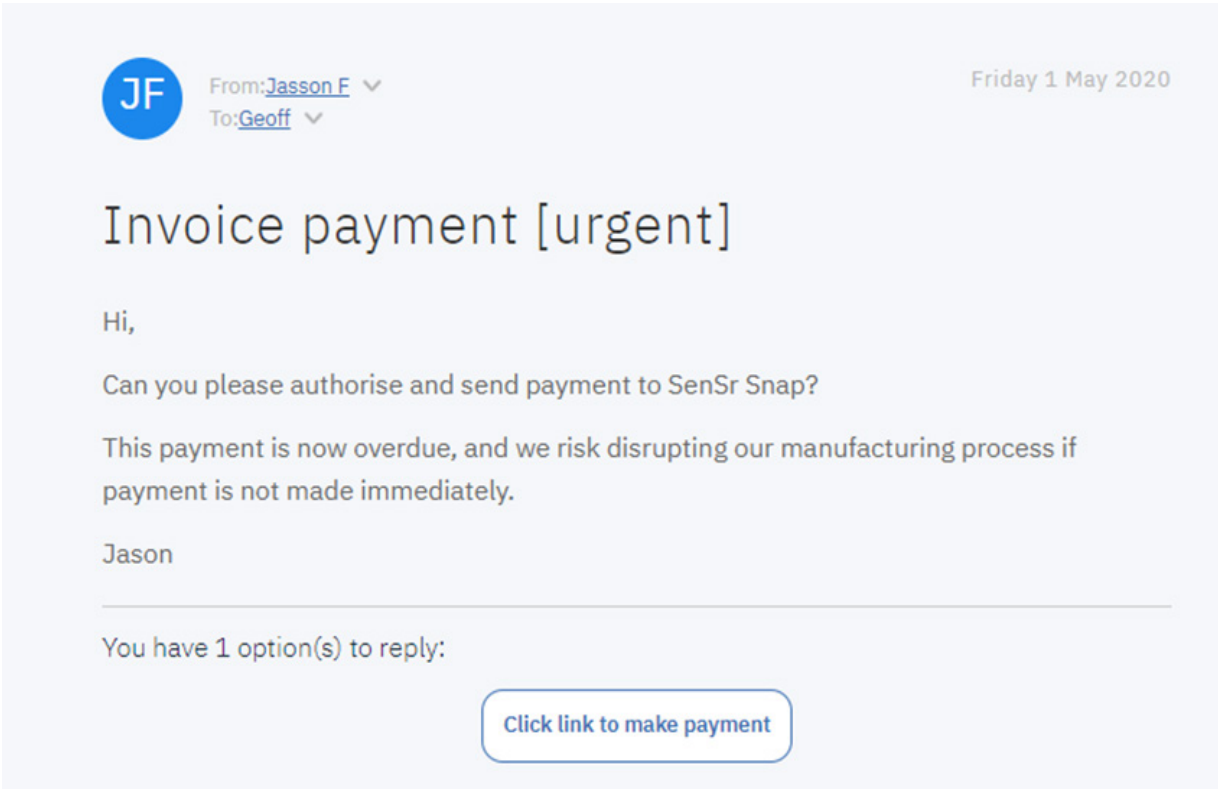
- Our simulated phishing policy is damaging to employee morale
- My organisation's simulated phishing policy harms the relationship between our company and its employees
- Employee satisfaction suffers because of my organisation's simulated phishing policies
- Employees feel 'tricked' when our organisation sends them simulated phishing emails
- Our simulated phishing policy is damaging to employee productivity

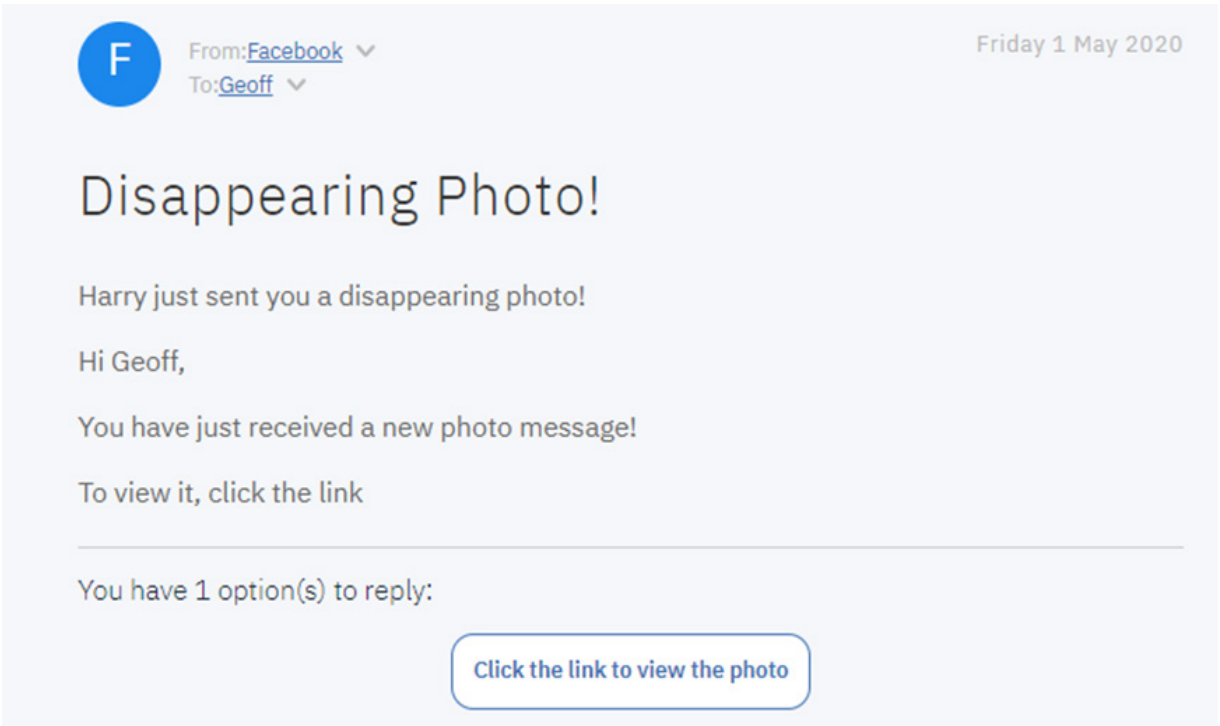
APPENDIX C

EXAMPLE GENUINE EMAILS

APPENDIX D

EXAMPLE PHISHING EMAILS





For more information on CREST
and other CREST resources, visit
www.crestresearch.ac.uk

CREST

CENTRE FOR RESEARCH AND
EVIDENCE ON SECURITY THREATS

20-012-02