Behavior change approaches for cyber security and the need for ethics.

Konstantinos Mersinas[1] and Maria Bada[2]

[1] Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK
[2] Queen Mary, University of London, Mile End Road, London, E1 4NS, UK

konstantinos.mersinas@rhul.ac.uk
m.bada@qmul.ac.uk

Abstract.

Humans are reportedly exploited as the main attack vector for security breaches. In order to minimize the susceptibility of humans to security attacks, it is not sufficient for individuals to just be aware, but they need to change their behavior as well. Such behavior change, that is, the modification of user behavior, can occur via targeted interventions, which are gradually being introduced in cyber security. In this paper, we identify and categorize the main approaches used to change user behavior and portray the main limitations of these approaches. Other fields, like health sciences, psychology and economics, have been traditionally more mature in ethics-related considerations. We suggest that although individual behavior change is increasingly being embraced by security practitioners and professionals, ethical aspects of the accompanied interventions are by large neglected in the field. We explore the ethical traditions of utilitarian, deontological and virtue ethics and their relations with security. We posit that ethical frameworks are needed for cyber behavior change interventions as a means to enhance security hygiene on both an individual and an organizational level.

Keywords: cyber security, behavior change, behavioral interventions, ethics.

Introduction

In the past two decades cyber behavior change (CBC) has been attracting attention, and theories, mainly from behavioral economics, have been employed to make choice architecture (the design by which choices are presented; Münscher et al., 2016) more effective in an interconnected world. CBC can be defined as both short- and long-term modification in the security behaviors of individuals. Certain user behaviors, e.g., people falling victims of phishing attacks, increase cyber risks which can be minimized by altering users' behaviors and habits.

Organizations run frequent cyber security awareness campaigns to minimize cyber risks at organizational and individual level. Cyber security awareness campaigns increase in amount and scope, however, security incident numbers are not reduced significantly (Alshaikh et al., 2019). Additionally, despite the increasing focus on awareness campaigns, these often fail to achieve active user engagement (Bada et al., 2015). A possible explanation for this lack of engagement might be the way cyber security mechanisms operate in workplaces, e.g., systems often allow users to be passive (Blythe, 2013), instead of promoting engagement. A CBC intervention would be a risk message tailored to users, aiming to increase awareness, but, most importantly, to encourage specific behaviors. This need for shaping secure behaviors makes CBC an invaluable tool for security professionals and organizations.

CBC interventions have only been relatively recently introduced in cyber security (Briggs et al., 2017; Coventry, 2014), and, to the best of our knowledge, there is a relative lack of scholarship on the relevant ethical considerations. Namely, despite CBC attracting increasing attention in the past few years, there is no ethical framework in the field to direct and set boundaries for such practices. In this paper, first, we describe the main behavior change approaches. Second, we describe three dominant ethical traditions and link them to CBC. Finally, we evaluate CBC approaches through the lens of the ethical traditions and aim in setting the bases for the creation of concrete ethical practices for implementing CBC.

The structure of our paper is the following. Section 2 discusses the existing approaches to behavior change. In particular, we focus on fear appeals, conceptual frameworks, the nudge and boost theories, nonconscious approaches, and incentives and disincentives. In Section 3, we link main ethics traditions with behavior change in security and in Section 4 we discuss the limitations and ethical considerations of behavioral interventions. We suggest that there is a need for ethics in cyber behavior change interventions and present concluding remarks in the last section.

Approaches to Cyber Behavior Change (CBC)

It is reported that up to 80% of security breaches are caused by "human error" as the underlying attack vector (Waldrop, 2016; Humaidi and Balakrishnan, 2015). Thus, it is critical that professionals and practitioners strengthen the so-called "human defenses" of users. The most common approach is security awareness training campaigns. Defenses, however, are not just dependent on awareness, as being simply aware does not guarantee appropriate, secure actions. Behavior is the key for avoiding human-originating breaches, and consequently, cyber behavior change is needed. We define behavior change as a modification of individual behavior achieved via some type of intervention. As expected, such interventions vary to the degree that they influence people's choices.

The economist Herbert Simon described the theory of bounded rationality, according to which, as humans, we are not fully rational agents and the optimality of our decisions is bounded by a number of factors. Namely, we have limited time, cognitive capacity and access only to a fraction of information for any given problem (Simon, 1972). This is not to diminish the influence of knowledge and understanding in optimizing decision-making, and more so in organizational contexts (Simon, 1991).

Beyond human error, other human-related factors have been studied in security. Namely, researchers try to overcome lack of understanding, negligence (Siponen and Vance, 2010) and apathy (Thomson and van Niekerk, 2012) for user non-compliance with policies. In this section we provide a selective review of the main approaches used to change user behavior.

Fear Appeals

Fear appeals have been a traditional tool in changing behavior. Witte defines fear appeals as "persuasive messages designed to scare people by describing the terrible things that will happen to them if they do not do what the message recommends" (Witte, 1992) or similarly, persuasive messages which convey the potential danger and harm if not adopting recommendations (Tannenbaum et al., 2015).

There is a number of aspects which can influence the effects of fear appeals on behavior change. Namely, the conveyed message itself, the type of behavior that is proposed and the characteristics of the audience (Tannenbaum et al., 2015). In more detail, relevant variables can be:

a) the level of conveyed fear;
b) the efficacy conveyed to the recipient, i.e., whether the solution is sufficient and/or effective;
c) the perceived efficacy of the individual, i.e., whether they are able to follow the recommended message;

d) the level of vulnerability and impact[1];
e) the style of the recommended action, i.e., whether the behavior is a one-off or repeated and whether it has a preventive or detective nature; and,
f) the characteristics (often demographic, e.g., age, gender) of the targeted individuals.

There is a number of theories focusing on different parts of the above aspects and variables. One of the first theories discussing how individuals react to fear appeals is fear-as-a-drive where drive is an unpleasant state that someone attempts to reduce (Dillard, 1994), ideally, by accepting the proposed message. Leventhal (1970), Sutton (1982) and Rogers (1975, 1983) consider fear arousal (emotional and physiological) as the decisive factor for behavior change, with higher levels of fear arousal being positively correlated with persuasion, but only if accompanied with high levels of perceived efficacy, i.e., beliefs that the conveyed solution (coping message) is effective (Rogers, 1975; Ruiter et al., 2014). And it has been empirically confirmed that combining threat appeals with solutions, increases the effect of the coping message (van Bavel et al., 2019). Other theories pose a dichotomy between linear models, i.e., models in which increased conveyed fear leads to increased behavior acceptance (Boster and Mongeau, 1984), or curvilinear, where high levels of fear are thought to cause the opposite effects to individuals (Sutton, 1982; Witte and Allen, 2000). Linearity seems to be supported by meta-analysis, along with the effectiveness of conveyed one-off behaviors compared to repeated actions, and evidence that fear appeals are more convincing for women compared to men (Tannenbaum et al., 2015).

Protection Motivation Theory (PMT)

PMT was developed with the intention to explore the effects of fear appeals on health-related behaviors of individuals (Rogers, 1975). PMT is based on four distinctive factors: the perceived severity of a threat, the perceived likelihood of the occurrence of such a threat, the efficacy of the recommended preventive behavior, and the way the individual perceives their self-efficacy in coping with the given threat (van Bavel et al., 2019). PMT was revised in 1983 by Rogers to include different ways to "initiate a coping process" (Milne et al., 2000, p. 108). The coping process is based on the identified *response efficacy* which is individuals' belief that they can deal with a threat effectively (Rogers, 1983). The theory is the first one to include the factor of self-efficacy for explaining human behavior in the light of a threat (Weinstein, 1993).

PMT has been applied in cyber security contexts; for example, by measuring employees' resulting compliance to security policies (Johnston and Warkentin, 2010; Siponen et al., 2014), and encouraging individuals to protect their systems,

[1] We use the cyber security terms *vulnerability* and *impact* here, although, in psychology and economics these are often termed as *susceptibility* and *severity*, although susceptibility is meant to have a strong personal relevance to the individual.

given that they know how to do so, but do not behave accordingly (Workman et al., 2008). Figure 1 depicts the process of PMT adjusted for a cyber security awareness training application (Mersinas and Chana, 2022). Namely, a message is conveyed along with the likelihood and the impact of a threat materializing. The recipient has a subjective perception of likelihood and impact, and also evaluates his or her own self-efficacy, along with the efficacy of the proposed solution. The result can be either acceptance of the message, via protection motivation, and consequently a change of behavior according to the recommendation, or a message rejection with inaction or an opposite action as a response.

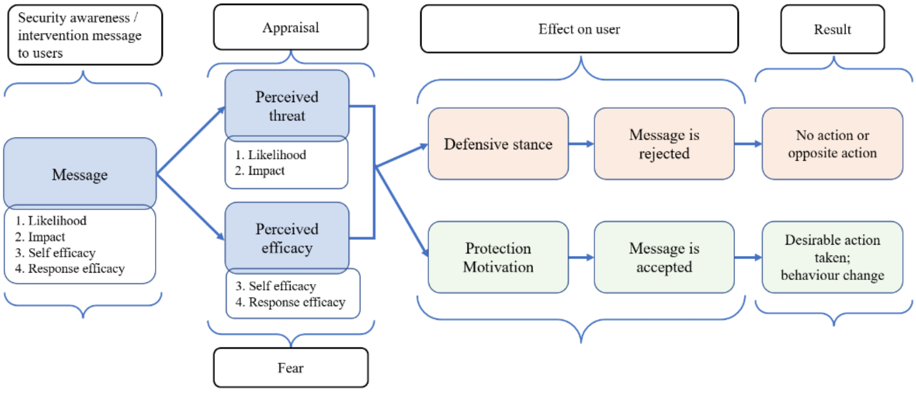

Figure 1: Fear appeals and the possible individual responses (Mersinas and Chana, 2022).

Structurally, PMT consists of two processes, the threat appraisal (threat message) and the coping appraisal (proposed solution). Van Bavel et al. (2019) use a coping message to inform users on how to deal with the threat and experimentally examine the effectiveness of informed coping messages with fear appeals for minimizing exposure to online risks. While both fear appraisals and coping appraisals contribute to protection motivation and secure behaviors, coping messages are shown to be comparatively more effective. Therefore, coping appraisals need to have a key role when designing behavioral interventions. Notably, PMT allows space for both environmental (observations, verbal persuasion) and intrapersonal factors (prior experience, personality traits) on the evaluation of threats and coping messages.

In security contexts, an individual's maladaptive, non-compliant-with-the-coping-appraisal evaluation can include the rewards of convenience, speed and simplicity for individuals, if these are perceived as larger than the risk. On the other hand, an individual's adaptive coping decision, might consider skills and capabilities (self-efficacy), how effective the solution is (response efficacy) in comparison with the costs of the recommended behavior. Thus, in both cases, usability of the proposed solution can play a role. The PMT model does not assume full rationality of individuals, but can equally work with, e.g., ecological rationality, i.e., interactions with the environment and usage of rules of thumb (Mersinas et al., 2019). However, in parallel, biases are potentially introduced in the model, as the evaluation of threats and coping messages are eventually subjective.

Theory of Reasoned Action (TRA) and Theory of Planned Behavior (TPB)

The Theory of Reasoned Action was proposed by the psychologists Fishbein and Ajzen in 1975 as a model to explain human behavior. The model has three main components: belief, attitude and intention, all of which produce a final behavior (Figure 1). In more detail, belief is an assigned probability to a cause-and-effect phenomenon. Attitude is the individual's evaluation of this phenomenon and it is a function of beliefs that lead to behavioral intention, which in turn is the likelihood of taking an action or following a specific behavior. The model was further amended to include subjective norms, i.e., normative beliefs (the evaluation of what others expect) and motivation to comply (the degree to which the individual wants to comply with other people's expectations); subjective norms are at the same hierarchical level as attitude (Figure 2).

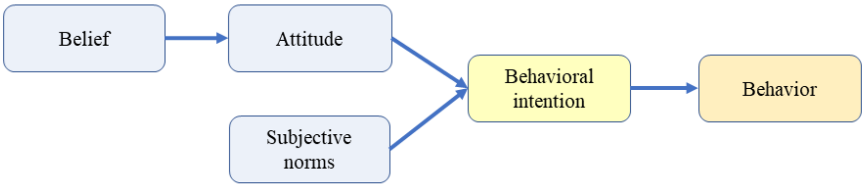

Figure 2: TRA model; adapted from Hale et al. (2002).

TPB is an expansion of the theory of reasoned action (TRA) and considers the additional component of perceived behavioral control as a factors which influences intention (Ajzen, 1991; Staats, 2004). This additional component is, as a construct, a synonym for self-efficacy, i.e., the level of control we believe we have over our behavior, but in practice it tends to be evaluated by how easy or difficult an action or behavior is perceived to be (Wallston, 2001). The addition of perceived behavioral control captures both relevant skills, e.g., digital literacy in security, and external conditions to be met, e.g., the existence of IT support or the existence of security mechanisms to be utilized.

Conceptual frameworks for behavior change

We identify two indicative examples in this category; the Fogg behavior model and the Hook model.

The Fogg model

The Fogg model (FBM) (Fogg, 2009) attempts to capture the components which need to coincide in order for a behavior to take place. The model proposes that the main factors for successful behavior change are a person's motivation, sufficient ability and effective triggers.

In more detail, Fogg's dichotomous variables for motivation can be pleasure and pain, hope and fear, and social acceptance and rejection. Ability can denote time, money, physical effort, brain cycles, social deviance, and non-routine. By brain cycles Fogg means the ability to think of a task while facing multiple everyday thoughts. Social deviance refers to acting contrary to the norm. Routine is easy because individuals are used to follow certain patterns, whereas non-routine actions reduce a person's ability to act as the simplicity of routine is removed. Simplicity, then, is a subjective factor because each person has a different concept of what is simple based on their background, skills, culture etc. The main elements of effective triggers are spark, facilitator, and signal. Fogg defines spark as an intervention design focusing on motivating the individual. Facilitator is a trigger for individuals who lack in ability even though their motivation is high. Signals are reminders and are suitable for individuals who have both the motivation and the ability to perform an action. Finally, motivation and ability are the deciding factors for the effective triggers.

The Hook model

The Hook model (Eyal, 2014) also includes triggers, along with actions, rewards, and investment and it is oriented towards habit formation. In particular, there is a trigger (usually external, but can be internal too) that causes the user to perform an action. Thus, the trigger is the event that actuates the action. The action needs to be relatively easy and is linked to an anticipated reward, which is then provided (at some point in time). So, the action is the expected or desirable behavior which is linked with the anticipation of the reward. Rewards can be of any kind, for example, material, social or personal (e.g., gratification via achievements). Rewards can be variable, with the intention of creating an increasingly 'addictive' feedback loop for the individual, so that they are motivated to repeat the action.

The main difference with Fogg's model is that the user has an opportunity to invest, e.g., in time and effort. This notion comes from a context of product design, for which the model was intended, but it can be adjusted in other applications too. The investment creates a connection with the "system or product" and thus, increases the chances that the user will repeat the process the next time the external trigger is provided.

Although FBM is simple and intuitive, its main issue is that it is intended as a conceptual framework. The constructs of motivation and ability are of a high level and less practical to implement. The Hook model is more focused on applications or specific features. However, the Hook model does not consider which triggers might be more effective and how to enhance them. Such models have been applied in social media and web applications and can be indeed powerful. But, as an example, user "addiction" associated with the feedback loop of a "like button" can be viewed as ethically manipulative because it relies on designed gratification via dopamine release in the brain, and thus, may reinforce the need of ethics.

Nudges and boosts

One of the main challenges of behavior change relates to choice architecture. Choice architecture refers to the multiple ways a choice can be presented to an individual and which can subtly direct them towards a specific choice. Choice architecture affects the individual's choice without always requiring their consent or their knowledge of that choice, which raises ethical considerations about the user's autonomy and have been even deemed anti-libertarian (Yeung, 2012, p. 133). For example, a widely used intervention appears in the form of nudging. Nudge theory holds that governments and organizations can direct individuals towards optimal decisions by slightly changing their behavior.

A significant topic of debate surrounding nudging pertains to the contradictions of the interventions. Indeed, nudges are intended to be libertarian, but are simultaneously paternalistic. Libertarian paternalism is defined as choice architecture stimulating choices believed to enhance the welfare of an individual, but at the same time maintains the individual's freedom to choose the deemed choice as a "suboptimal" course of action (Thaler and Sunstein, 2003). For example, opt-out policies are a form of libertarian paternalism because they provide the user with the choice to not choose a default option. Opt-out policies can be effective because of the additional steps (the so-called transaction costs) an individual has to take in order to change the situation or default option. Therefore, the success of such policies relates to human inertia – and the power of defaults – rather than persuasion.

Caraban et al. (2019) identify 23 ways to nudge in human-computer interaction and link cognitive biases with the mechanisms of nudging in cyber security. These can be categorized into nudges targeting either the reflective or the automatic mind, and nudges which are transparent or non-transparent. Thaler and Sunstein, the creators of nudge theory, advocate a strong transparency via visibility and monitoring. This approach can avoid manipulative behavioral architectures, i.e., designs where targeted individuals do not know the intentions behind an intervention or even realize its implementation. However, strong transparency can be restrictive, even for policies beneficial for the public (Hansen and Jespersen, 2013). Thus, it is argued that even if manipulation can be avoided, full or strong transparency might undermine the well-intended and beneficial outcomes of a nudge.

Another approach to transparency is based on the reflective or automatic functions of the mind. *Reflective* indicates the part of the mind that is processing information slowly, effortfully and intentionally (called System 2). We can say that it is the controlled part of the mind. The automatic mind, on the other hand, is processing information fast and without active deliberation (System 1). When a nudge is stimulating the reflective part of the mind, the process is transparent because the individual can process the information and act accordingly. When the nudge targets the automatic mind, the process is non-transparent because individuals react intuitively to the nudge.

Although, in general, the individuals' freedom of choice is maintained, an ethical concern is that some nudges lack transparency, because the process leading to the nudge "may be far more secretive". In particular, Baldwin identifies three degrees of nudging, from simple information that maintains full autonomy of the target, to building on volitional limitations of individuals (e.g., by using defaults and opt-out policies), to the third degree that interfere with autonomy and reflection by utilizing salience, framing and affect (Baldwin, 2014). The effect of the aforementioned opt-out mechanisms, has been explored by various researchers, and they can be considered as cunning, although, "not all opt-out mechanisms raise ethical questions" (Caraban et al., 2019). However, nudges can lie between coercion (full control of the influencer) and persuasion (no control over the influence) and it is debatable as to whether they maintain freedom of choice (Saghai, 2014). Indeed, for a fully autonomous choice individuals need to be rationally persuaded, since "only rational persuasion fully respects the sovereignty of the individual over his or her own choices"; however, they highlight that emotions play a role in rational persuasion (Hausman and Welch, 2010, p. 135).

The need to rationally persuade an individual has given rise to boost theory (BT), a design aiming to improve people's decision-making by helping individuals reach their highest possible capacity to achieve goals. BT targets competences and individuals' agency instead of immediate behavior (Hertwig and Grüne-Yanoff, 2017). The main difference with nudge theory is that BT allows individuals to reflect on the decision, while nudges change behavior mostly through the choice architecture. The individual is provided with the optimal course of action and is expected to decide actively and transparently (Grüne-Yanoff and Hertwig, 2016; Hertwig and Grüne-Yanoff, 2017). Boosts are an enhanced form of nudges, as they are focused on longer-term behavior change. To achieve long-term results subjects need to have access to all the intervention parameters.

Apart from long-term behavior change, boosts can be useful in short-term changes. Short-term boosts encourage the development of a competence in a specific context, and they resemble a so-called "educative nudge", aiming "to overcome or correct behavioral biases by promoting learning" (Sunstein, 2016). Long-term boosts aim to render a competence readily used at will and in various contexts. The ideal result of a long-term boost is permanent behavior change. Long-term boosts are categorized depending on their goals. Namely, Hertwing and Grüne-Yanoff (2017) distinguish long-term boosts into risk literacy, uncertainty management and motivational boosts. Risk literacy boosts aim to render people able to comprehend statistical information for a wide range of domains. Uncertainty management boosts focus on the ability of the subject to assess a situation in uncertain conditions and motivational boosts motivate subjects to act while maintaining their autonomy.

Boosts require the subject's informed consent and value autonomy, however, they are sometimes criticized for inducing significant costs and effort to intervention recipients; and indeed BT interventions require time and cognitive resources. And even in the cases that boosts are low-cost for subjects, they can be challenging for policy makers. Policy makers bear high cost of boosts due to the required complexity of the interventions. The main advantage of boosts is the formation of habits, which require active effort for their initial formation. The potential usefulness of boosts can be inferred given that, reportedly, up to half of our actions and decisions are the byproduct of habits (Wood et al., 2002).

Nudges can be considered to threaten the autonomy of individuals since policymakers do not know people's true interests. This view is based on the definition of autonomy by John Stuart Mill, that autonomy is the ability of individuals to decide their own interests and make choices based on these interests (White, 2013). More specifically, nudges can be seen to violate individuals' freedom of choice and autonomy, especially if the influencer's or policy-maker's intentions are unclear or if there is no recipient consent for the intervention (Hausman and Welch, 2010; Wilkinson, 2013). On the other hand, boosts require individuals to respond to interventions in a motivated, reflective fashion, and thus, avoid considerations on limiting individuals' autonomy.

Nonconscious behavioral approaches

Nonconscious behavior[2] is any behavior which is not processed consciously by the brain. Thus, the result of nonconscious interventions is not intended by the individual performing the associated action. Such interventions target the automatic part of the brain (System 1) rather than the reflective (System 2). System 1 is uncontrolled, effortless, associative, fast, nonconscious and skilled (Thaler and Sunstein, 2008). Nonconscious influence is often triggered by subliminal stimuli, e.g., usually visual or auditory stimuli that individuals are not consciously aware of and which are either hidden or used in a way to *prime* individuals (i.e., use stimuli that influence subsequent actions).

Despite the disparity between conscious and nonconscious mental processes, nonconscious activity has been found to be important and beneficial in decision-making, e.g., by leading to fast and optimal decisions by experts, after years of accumulated experience (Wegner, 2002), indicating that the value of underlying nonconscious processes should not be ignored.

Nonconscious approaches have not been explored in depth in cyber security, apart from certain nudges targeting the automatic part of the brain (such as opt-out policies, to an extent) (Caraban et al.,2019). But, studies targeting nonconscious behavior are appearing in digital behavior change interventions, in particular with the aim to avoid decision reliance on motivation and ability, since these are volatile (Adams et al., 2015). Interventions based only on conscious cognition or only on automatic nonconscious processes are considered as less effective in forming long-term behavior change. For that reason, (digital) interventions which utilize both the automatic and the reflective parts of the brain are proposed in the literature (Pinder et al., 2018).

From an ethical perspective, in nonconscious interventions subjects are by definition manipulated and are unaware of this manipulation. Therefore, nonconscious approaches can be viewed as undermining the autonomy of the individual and, thus, raise ethical considerations. This aspect is similar to a category of nudges that are not controlled by the targeted individual, and, thus, undermine freedom of choice (Saghai, 2014). It should be noted, however, that intervention recipients' responses are not necessarily purely automatic or nonconscious. There is evidence

[2] We use the term *nonconscious* to cover both the term *subconscious* (processes not in focal awareness) and *unconscious* (deeper mental processes).

that individuals' conscious volitional decisions are influenced by nonconscious environmental factors (Parkinson and Haggard, 2014).

Additionally, nonconscious interventions also lack transparency by definition and, thus, may lead to suspicion or mistrust in various settings, resulting in more harm than benefit. And in fact, similarly to the discussion on nudges, a lack of transparency can allow for the goals behind the intervention to be questioned. In that sense, interventions directly communicated to users and accompanied by persuasion techniques, overcome these considerations.

Incentives and disincentives

An incentive or reward (or a praise), in the broader sense, can be anything that motivates an action and can be intrinsic or extrinsic. A disincentive or a punishment (or a blame), symmetrically, is anything that withholds or removes a reward or applies some 'painful' stimulation. There is an ongoing debate on the relative effectiveness of either approach. But, research evidence indicate that rewards work better for motivating action, whereas punishment is more efficient for deterring individuals from taking an action; this finding is based on how our brain has adapted to our environment (Guitart-Masip et al., 2014).

Incentives are used by organizations as the first step to policy compliance, including security policy compliance. The reasoning behind punishments (or sanctions) is to an extent based on the criminological General Deterrence Theory (GDT) which has been used widely – including the security field – to examine whether punishment is an effective means to change behavior. According to GDT an individual chooses to obey or break "the rules" based on rationally analyzing potential consequences (Andenaes, 1974), and certainty, quickness, and severity of punishment influence the decision (Theoharidou et al., 2005).

The predictive power of rewards and punishments to encourage security policy compliance is found to be weak, especially when these are imposed through specific guidelines (e.g., specific policies like related to anti-virus) (Cram et al., 2017). An issue with the reasoning of GDT also, is that fully rational agents are not necessarily observed in real-world scenarios, especially with regards to probability estimations (Kahneman, 2011), i.e., the *certainty* aspect of GDT.

Cram et al. (2017) point out the need for new research to understand what type of incentives would be the most effective for organizations to implement in order to balance effects on the organization and the individual, in a security context. Goel et al. (2020) provided financial rewards to a small sample of employees and the result was an improvement in hygienic security behavior, namely, stronger passwords. Once the program was over, however, employees showed signs of increasing non-compliance due to the limited temporal effects of extrinsic incentivization.

Outside cyber security, experiments examining extrinsic (financial) incentivization effectiveness have been conducted in health sciences, indicating that participants fall back to their previous behaviors, long-term (Carrera, 2018). Financial incentives are impactful for the duration of their implementation, but they do not lead to habit formation. And due to the failure of forming habits, employees potentially become non-compliant once the reward intervention ends.

Punishment, as a means to change behavior, can also be viewed by employees as an unjust measure of organizational control, and the extent of such a perception is also culture-dependent. Namely, in the more individualistic western cultures (Nisbett, 2004), such perceptions of control might be more prominent. Punishment, on the other hand, is perceived as necessary sometimes to ensure the smooth operation of an organization. However, punishment, if utilized, should be balanced with incentivization by rewarding conformity in order to build trust relationships within an organization (van der Akker et al., 2009). It should be noted that punishment in an organizational security context, mostly refers to sanctions (i.e., assignment of liability) rather than penalties or monetary fines.

On a neurological level, rewards can have similar effects to punishment, if their provision is halted. Shabel et al. (2019) experimentally evaluate stress effects on the lateral habenula, the part of the brain responsible for decision-making. Results indicate that stress causes the brain to react with punishment signals when a reward is withdrawn, thus, equating a lack of rewards with punishment. That is, the signal for reward omission is the same as that for punishment. Security practitioners who wish to use incentives could consider mixed methods, i.e., both rewards and punishment, but with a long-term orientation.

Behavior change ethics

We utilize three main ethics traditions and draw links with cyber security and interventions for behavior change. Ethics and philosophical approaches have been used in Information and Communication Technologies (ICT), but they are usually focused on product or policy design (Siponen and Iivari, 2006; Brey, 2015; Bednar and Spiekermann-Hoff, 2020) and not as a lens to identify how users need to be treated in a field which has a polemic (attacker-defender) nature and thus, specific narratives are conveyed to users, which can influence their attitudes. We explore the applicability of three main dominant philosophical and ethical traditions, dominant at least in the West (Bednar and Spiekermann-Hoff, 2020); namely, virtue ethics, deontological ethics and utilitarian ethics.

Utilitarian ethics

Utilitarian ethics are a special case of a moral view called *consequentialism* and they focus on the common good or the “overall consequences”. The individual is expected to act in fashions which can be deemed “good” if they progress the greater good. In this tradition, it is the outcomes that matter, i.e., there is a focus on ends, not how they are achieved. *Utility*, the term mostly used in economics, is the equivalent of happiness and well-being, or anything of value, and is what needs to be maximized under utilitarianism (Bentham, 1876; Mill, 1859). Therefore, the tradition is based on rational decisions regarding the overall good and assumes the use of a cost-benefit analysis for decision-making which weighs whether the utility of the most people is maximized in comparison to one or a few.

One aspect of this approach is that a utilitarianist will always choose "society" over the individual and this might create a series of issues relating to individual rights. In particular, if the individual is of secondary importance compared to the group, individual rights like privacy rights and control over personal information, might be undermined for the sake of the majority, or the greater good of an organization. Assuming that there is no regulation violated, utilitarian ethics are in line with most business practices, considering the organization as the analogous of "society". Thus, practices as merges, departmental restructures, human resource management, relocation and employee firing are considered ethical in these terms.

In more specific security settings, employees of an organization with utilitarian ethics in place must follow organizational policies. Thus, compliance with security policies is justified under the goal of an overall protection of the organization. However, this does not exclude non-compliant behaviors. In the case that employees are not to comply, they should do so by having in mind the benefit of the majority of the organization's members of staff or the overall good. An immediate consideration here is the subjectivity of the justification for non-compliance and its potential confirmation only on hindsight.

In this ethical tradition, beyond maximizing the overall utility, avoiding harm for the majority is equally a main goal of individual actions. This does not necessarily exclude approaches like sanctions, or "blaming and shaming" of individuals. Consider the scenario of an internally executed phishing campaign. Suppose that senior management decides to publicly shame employees who were tricked by the phishing emails. If this action is considered as a means for the overall good of the organization, then it is in line with utilitarian ethics. There is however, one issue with this reasoning; namely, it is difficult to measure the effects on the overall utility. It might be the case that more employees comply out of fear, thus, the desired outcome is achieved (setting aside the discussed limitations of fear appeals, for the sake of the argument here). But equally, many may be disillusioned with senior management and lose any sense of trust, a result which might defy the whole point. Thus, in practice, it is hard to justify this approach, especially, if the well-being of specific individuals is directly and measurably affected.

Deontological ethics

The second tradition is deontological ethics, a thought system promoting the single ideal of acting in ways that an individual wants the whole society to follow. It is a human oriented system and promotes collective thinking and a "universal moral obligation". The term deontology derives from the Greek words *deon* and *logos*, meaning duty (or necessity) and reason, respectively, and indeed, universalizable rules of conduct (or morality) are based on reason in deontology. This tradition is attributed to Immanuel Kant and his Categorical Imperative which states: "Act only according to that maxim by which you can at the same time will that it should become a universal law" (Kant, 1998, p. 422). In contrast to utilitarianism, here it is the nature of an act, rather than the outcome that matters.

Every individual must follow the universal rules established by the deontological system of thought. Morality and ethics become an obligation, however, there is a reported misconception about deontology that individuals cannot have incentives or freedom of choice (Van Staveren, 2007). In fact, individuals are assumed to voluntarily comply with the accepted moral rules, in line with Kant's imperative. For example, consider the scenarios where employees accept a decision by senior management, or citizens comply with election results, or tax payers accept to pay additional contributions during an economic crisis. All these behaviors can contradict individuals' utility maximization, but they have some accepted underlying morality, therefore, individuals voluntarily agree to behave accordingly. This underlying morality corresponds to the *belief* and *attitude* stages of the fear appeal theories TRA and TPB.

Since deontology has the notion of 'ought' to do something, it is linked with individuals' moral obligation, because individuals have a sense of duty in their ethical actions. Consequently, and to the extent that deontology might influence individuals to act in certain ways by inducing a feeling of guilt (Cronan and Al-Rafee, 2008), there can be well-being considerations in organizational settings. This issue is similar to the individualized effects of utilitarian ethics already discussed.

A possible conflict of deontology with traditional security risk management can be the fact that actions are inherently either right or wrong, independently of their potential impact. Impact (and likelihood) is an established way of thinking in information security and a core notion of fear appeals, for that matter. The notion however, is not a core one in deontology, but it could be introduced under the deontological ethics angle of rationality. On the other hand, since reason plays a role in deontology, individuals are not expected to comply with policies without justification, it is just a matter of establishing that it is, e.g., a universal and accepted rule to protect organizational information assets, without necessarily focusing on the impact of, say, regulatory fines.

Finally, organizationally, deontology presupposes some form of authority and, thus, might be in line with top-down approaches to security and typical business hierarchies. Moreover, behavioral interventions based on deontology can be more practical in their implementation within the commonly established organizational hierarchical structures.

Virtue ethics

Virtue ethics is a branch of ethics founded by Aristotle, according to whom, humans strive for *eudaimonia,* that is for happiness and flourishing. Eudaimonia is the highest of the goods and it is worth pursuing it for its own sake (Aristotle, 1980). Virtue ethics are applied contextually, in contrast to deontological ethics, which, as we have seen, attempt to define universal rules. This context-dependency and the voluntary nature of decisions create a component of *responsibility* (Van Staveren, 2007).

Virtue ethics have an individualistic angle which allow for voluntary action. For example, this can mean users having the freedom to follow or ignore security policies, exceptionally (Siponen and Iivari, 2006) or that employees can act based on their own judgment. However, virtue ethics cannot be considered as purely individualistic as humans are considered as social beings who act in relation to others (Van Staveren, 2007). The other strong characteristic of this ethical tradition is that virtues are self-sufficient and thus, people follow "good" actions for their own sake, not as a means to other goals (Aristotle, 1980).

In the latter case, actions can be considered acceptable if they satisfy three criteria: be just, honest and courageous (Siponen and Iivari, 2006, p. 454). In that sense, employees need to have a sentiment of justice, their intentions to be guided by honesty, and be courageous and take the lead in an autonomous manner without needing supervision. These notions are highlighted by Aristotle as practical and concrete behaviors, e.g., in daily interactions and not as abstract rules (Van Staveren, 2007). Thus, virtue ethics can work under the assumption that users are educated in security, and have a self-efficacy level and behavioral control over potential actions, which would allow them to judge situations independently.

The contextual role of responsibility can be pivotal in a security setting. Namely, security and risk perceptions of employees who have responsibility and involvement with security mechanisms and processes is found to be more positive than that of individuals who are unrelated to these mechanisms and processes (Durojaiye et al., 2020).

Moreover, virtue ethics attempt to establish a middle way between reason and emotion (often termed intuition), therefore they are in contrast with approaches like the Hook model, which target emotional reactions. The self-sufficiency of virtues might be in contrast with behavioral interventions and models which utilize rewards and incentivization. Instead, it implies that, e.g., compliance messages to individuals are self-evident. Such an attribute might be incompatible with most organizations, since it is hard to imagine, e.g., security policy compliance, because "it is a good thing". Thus, we see a possible mismatch with cyber security.

In parallel, there are arguments that emotions do have a role in ethical reasoning by recognizing human limitation and vulnerabilities (Nussbaum, 2001). In that sense, approaches like fear appeals which utilize emotions can be in line with virtue ethics, and in particular, if we expand Nussbaum's arguments, as a way for individuals to reflect on their own vulnerability and their role in the security context, via fear of a conveyed message about a potential security incident. Virtue ethics thus, relate to character; and therefore, the tradition is well-suited for behavior change interventions, to the extent that these take into account individual characteristics, like personality traits. This is not to say that all behavioral interventions take this approach, but it is one that is most promising given the aforementioned variables of fear appeals and the Fogg model, in particular.

From the social angle of virtue ethics and the importance of agent relationships and a voluntary commitment to shared values (Van Staveren, 2007) we derive that they require established security practices. In particular, virtue ethics might work better with established (or developing) security cultures, ideally positive ones. Thus, virtue ethics might be more appropriate in relatively mature security environments as they can reinforce behaviors via interactions and norms.

Discussion on the limitations and ethical considerations of behavioral interventions

In this section, we highlight limitations and ethical considerations which span across CBC approaches, that is, fear appeals, conceptual frameworks, nudges and boosts, nonconscious interventions, and rewards and sanctions. Some considerations are uniquely associated with a CBC approach, while others apply to more than one approaches. The identified points of consideration, along with the key notions within the ethics traditions, indicate a structure for the development of an ethical framework. That is, the limitations and considerations, along with the key notions from the ethics traditions, are meant as learnings and a basis to shape specific ethical frameworks in future research. More specifically, this basis is comprised of autonomy, social responsibility, the common benefit, individual rights, non-harm, transparency, and a justification of the interventions.

Fear appeals limitations. A number of ethical considerations surround the fear appeals literature outside security. Indicatively, the violation of autonomy in health-promoting strategies (Tengland, 2012), the causing of distress to the targeted individuals in anti-smoking campaigns (Hastings et al., 2004), and the causing of anxiety and other negative emotions in the context of emotion-arousing ads (Hyman and Tansey, 1990) are all ethics-related reported issues. But also, specifically in security, researchers have proposed ways to enhance behavior change models. For example, Jonston et al. (2015) suggest that fear appeals and PMT models are inadequate for security and propose the incorporation of personal relevance in the conveyed messages as a means to enhance compliance. Indeed, research findings indicate that both environmental and individual factors need to be accounted for, along with behavioral interventions, e.g., a lack of time, knowledge or skills can affect self-efficacy levels (Reid and van Niekerk, 2016).

Neglected influencing factors. Thus, the main limitation of fear appeals is that they do not consider behaviors which necessitate a wide range of additional factors like skillsets, opportunities and context. TPB has some useful features, namely, it attempts to capture situations where individuals have reduced control. For example, it considers the situation where, despite being motivated, an individual might fail to perform and action if the required environmental conditions are not available. But a significant limitation of for both TRA and TPB is that volition and conscious will are presumed; the difference between the two theories is that TRA assumes full volition, whereas TPB introduces the behavioral (internal and external) control which can hinder full volition. That is, individuals utilize beliefs, evaluate them via attitudes and, thus, consciously form intentions. Thus, there is a consideration for the so-called intention-behavior gap in TPB, in line with the observation that the causal relationship between intentions and behavior is not straightforward. Interestingly, the presumption of this relationship is compatible with the assumptions of most ethical traditions.

Short-term, non-habitual effects. While fear appeals can be successful in changing behaviors, they do not necessarily form habits. Habitual conduct is largely unaffected by intentions only, whereas small and gradual behaviors can form habits. Notably, the formation of habits is associated with long-term effectiveness of interventions and the goal of a positive security culture. Extrinsic incentivization alone is shown to be ineffective in this direction, but a combination of components might be useful; namely, voluntary action, engagement and responsibility within social interactions are in line with virtue ethics and might shape a security culture. The "societal utility" of utilitarianism might be in line with organizational goals and an overall security culture at first glance, however, a positive security culture needs to be built equally on individualism and, thus, the individualistic nature of virtue ethics and deontology might be a better fit. And we have discussed that the security culture maturity might also be a contextual factor, e.g., a positive security culture with an established notion of security as a "good" might be in line with virtue ethics.

Non-plurality of choices. Behavioral interventions can be seen to violate users' autonomy because users are led to follow a specific route of action, the so-called coping appraisal, dictated by, e.g., security professionals. The problem with this attitude is the creation of a paternalistic approach, i.e., dictating specific solutions, since a specific action might be imposed on users leading to choice restriction. That is, a lack of alternatives, might take place, since users are usually presented with two choices, the one being "optimal" (via the provided coping appraisal), whereas the other "dangerous" or "irresponsible". The "optimal" option is presented as the only logical and legitimate choice provided by intervention designers.

Distress. Fear appeals can also cause distress and those exposed to fear may be unable to act on the relevant advice. "Fear, Uncertainty and Doubt" (FUD) has been criticized as an unethical and unhelpful practice because many of the factoids shared through FUD aim in creating an unpleasant atmosphere for the recipient along with the elicitation of fear (Florêncio et al., 2014). Beyond the ethical issue raised by FUD and although it is not an unusual appeal in cyber security, its effectiveness is not clear. Namely, security breach reports and headlines often utilize FUD to convey messages possibly leading to fatigue. Another factor which potentially diminishes the effectiveness of FUD is cognitive biases, like the overconfidence bias, also called the "it will not happen to me" bias. Thus, fear-based approaches can be seen as an attempt to inflate perceived risks, but cognitive biases may work in the opposite direction, affecting the objective estimation of the threat likelihood.

Opposite effects. The use of fear in behavioral interventions might be ineffective as, for example, in certain occasions people tend to respond to fear with humor, and, thus, undermining the effectiveness of the interventions. This finding is observed in Twitter posts (Abril et al., 2017) an online platform which might have similarities with an environment that conveys security messages to employees. Humor responses to fear are called fear control responses and are a psychologically legitimate way of coping with fear and unpleasant feelings (Martin, 2010), but their existence confirms the ethically questionable instigation of unpleasant feelings.

Well-being risks. The well-being of individuals, in the broader sense, is a main concern as fear and disincentives can induce unpleasant emotions, either directly or via peer pressure. The same considerations hold for the application of rewards and sanctions; i.e., these can affect the well-being of employees. Additionally, disincentives as responses to user behavior can affect security culture. Namely, the demonization of those behaving insecurely can have a negative impact on long-term security behaviors by, e.g., targeting or blaming individuals or creating stereotypes (Renaud and Dupuis, 2019).

Manipulation. Certain nudges, as well as nonconscious approaches raise concerns for depriving autonomy and manipulating individuals. Since nudge theory works on the mantra that people can choose to act upon the nudge or not, most common behavior change theories assume that individuals form intentions by processing information via their reflective (System 2) rather than the automatic mind (System 1). The effects and ethics of subliminal messages have caused concerns in our societies decades ago, especially in advertising contexts, but seem largely neglected, with some occasional exceptions. Indicatively, in US politics, George W. Bush's campaign portrayed images of Al Gore along with the word "RATS" repeatedly flashed for fractions of a seconds on the screen (BBC, 2000). But, by large, although experimental research indicates the influence of subliminal messages on individuals, there does not seem to be a broader concern. Maybe the prevalence of cyber security across societal functions can refocus discussions on nonconscious messages.

Non-specificity. Conceptual models of behavior change might be useful for educational purposes, but their generic nature reduce their practical value. Therefore, they can guide behavioral interventions at high-level, but lack the specificity needed in industry implementations. For the Hook model in particular, ethical considerations can be raised on the mechanisms underlying the provided rewards. Namely, rewarding feedback loops with unexpected but desirable rewards are shown to be associated with surges of the neurotransmitter dopamine in the brain. Dopamine suppresses reasoning and triggers behavior based on desire. Bypassing System 1 thinking, is a consideration of similar nature to subliminal messages and could undermine the autonomy of individuals.

Our aim was to showcase that behavior change approaches entail ethical considerations for their application. In our exploration, we ultimately aimed in showing that creating ethical frameworks for cyber security interventions is not a straightforward endeavor, but requires a synthesis of various components, as ethical traditions might need to be, first consulted, then, adapted and expanded to serve the security field and the contextual characteristics of the organization.

Conclusion

The way to utilize behavioral interventions in cyber security in an ethical fashion has not been fully explored yet. In this paper, we first highlight that such interventions are complex and no approach is free from limitations in its implementation. Second, we portray the ethical considerations of these interventions, advocating that ethics need to be introduced in security research and security awareness training practice. We present the ethical issues and the limitations surrounding behavior change approaches and posit that ethical frameworks need to be considered for utilizing the increasingly recognized need for behavioral interventions in security. The security field does not have a tradition of such approaches and therefore, we argue that a set of widely accepted principles, synthesized from well-studied ethical traditions is needed as a guide for professionals, practitioners and behavioral intervention designers.

We posit that a discussion on ethical behavioral interventions can be initiated in security and that a synthesis of the aforementioned ethical traditions can be adapted to the requirements of the security field and the organizational environments. In our analysis, a number of components are identified as possible building blocks for ethical frameworks for changing security behaviors. Namely, user independence and autonomy, social responsibility, the appropriate use of rewards or sanctions, and the transparency of interventions. Additionally, through the exploration of ethical traditions we portray that individual rights need to be protected and balanced with the greater organizational benefit.

Interdisciplinary research would further contribute in this area via, at least, two directions. First, by studying each of the aforementioned components of autonomy, responsibility, rewards and sanctions, transparency and individual rights in specific security contexts with different requirements, to identify how well they 'fit' real-world settings. Second, by analyzing and/or formalizing the ethical traditions and contrasting them to organizational cultures and hierarchies, to map characteristics of the traditions with real-world modi operandi.

Finally, we draw links between behavioral interventions and ethical traditions on the one hand, and security culture on the other. We hypothesize that different groups might have preferences for different ethical frameworks; for example, a perceptional dichotomy between policy makers and end users could exist. Thus, in future research we aim in examining perceptions and the feedback of security professionals and users, to crystalize such an ethical framework for behavioral interventions in cyber security.

References

Abril, E. P. and Szczypka G. and Emery S. L., 2017. LMFAO! Humor as a Response to Fear: Decomposing Fear Control within the Extended Parallel Process Model. *Journal of Broadcast Electronic Media*, 61(1), pp. 126-143.

Adams, A. and Sasse, M. A., 1999. Users are not the enemy. *Communications of ACM*, 42, 12 (Dec. 1999), pp. 40–46.

Adams, A. T., Costa, J., Jung, M. F. and Choudhury, T. 2015. Mindless Computing: Designing Technologies to Subtly Influence Behavior. *UbiComp '15, ACM*, 719–730.

Ajzen, I., 1991. Theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, pp. 179-211.

Alshaikh, M., Humza, N., Atif, A. and Maynard, S. B., 2019. Toward Sustainable Behavior Change: An Approach for Cyber Security Education Training and Awareness. In Proceedings of the *27th European Conference on Information Systems (ECIS)*, Stockholm and Uppsala, Sweden.

Andenaes, J., 1974. *Punishment and deterrence*. Ann Arbor.

Ariely, D., 2008. Predictably irrational: The hidden forces that shape our decisions. New York.

Aristotle, 1980. *Nichomachean Ethics*, trans. D. Ross and rev. J.L. Ackrill & J.O. Urmson. Oxford University Press.

Armitage, C. J. and Conner, M., 2000. Social cognition models and health behavior: A structured review, *Psychology and Health*, 15:2, 173-189.

Bada, M., Sasse, A. and Nurse J. R. C., 2015. Cyber Security Awareness Campaigns: Why do they fail to change behavior?. *International Conference on Cyber Security for Sustainable Society*.

Baldwin, R., 2014. From Regulation to Behavior Change: Giving Nudge the Third Degree. *The Modern Law Review*, 77(6), pp. 831-857.

Barford, L., 2019, November. Contemporary virtue ethics and the engineers of autonomous systems. In *2019 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1-7). IEEE.

BBC News, 13 September 2000. *RATS ad: Subliminal conspiracy?* http://news.bbc.co.uk/1/hi/in_depth/americas/2000/us_elections/election_news/923335.stm (Accessed: 25/01/2023).

Bednar, K. and Spiekermann-Hoff, S., 2020. The power to design: exploring Utilitarianism, Deontology and Virtue Ethics in three technology case studies. ETHICOMP 2020, p.396.

Bentham, J., 1876. An introduction to the principles of morals and legislation. Clarendon Press, Oxford.

Blythe, J. M., 2013. Cyber Security in the Workplace: Understanding and Promoting Behavior Change. *Proceedings of CHI* 2013. Doctoral Consortium, Trento, September 16th 2013, pp. 92-101.

Boster, F.J. and Mongeau, P., 1984. Fear-arousing persuasive messages. *Annals of the International communication Association*, *8*(1), pp.330-375.

Brey, P., 2015. Design for the value of human well-being. Handbook of ethics, values, and technological design: Sources, theory, values and application domains, pp.365-382.

Briggs, P., Jeske, D., Coventry, L. 2017, Behavior change interventions for cybersecurity. In: Little, L., Sillence, E., Joinson, A. (eds.). *Behavior Change Research and Theory*. Elvesier, Amsterdam, pp. 115-136.

Camerer, C., 2003. Behavioral game theory: experiments in strategic interaction. New York.

Camerer, C. F., 2004. Prospect theory in the wild: Evidence from the field. In C. F. Camerer – G. Loewenstein – M. Rabin (eds.). *Advances in behavioral economics*. Princeton and Oxford, pp. 148-161.

Caplin, A., 2003. Fear as a Policy Instrument. *Time and Decision*, pp. 441-458.

Caraban, A., Karapanos, E., Gonçalves, D., Campos, P., 2019. 23 ways to nudge: A review of Technology-Mediated Nudging in Human-Computer Interaction. *CHI 2019*.

Carpenter, P. and Roer, K., 2022. The Security Culture Playbook: An Executive Guide To Reducing Risk and Developing Your Human Defense Layer. John Wiley & Sons.

Carrera, M., Royer, H., Stehr, M., Syndor, J., 2018. Can Financial Incentives Help People Trying to Establish New Habits? Experimental Evidence with New Gym Members. *Journal of Health Economics*, 58, pp. 202-214.

Conner, M. and Sparks, P., 1996. The theory of planned behavior and health behaviors. In: M. Conner and P. Norman (Eds.). *Predicting health behavior* (pp. 121-162). Buckingham, UK.

Coventry, L., Briggs, P., Jeske, D., & van Moorsel, A., 2014. SCENE: A Structured Means for Creating and Evaluating Behavioral Nudges in a Cyber Security Environment. In A. Marcus (Ed.), Design, User Experience, and Usability. *Theories, Methods, and Tools for Designing the User Experience*, pp. 229–239.

Cram, W. A., Proudfoot, J., and D'Arcy, J. 2017. Seeing the forest and the trees: A meta-analysis of information security policy compliance literature. Proceedings of the *50th Hawaii International Conference on System Sciences,* (2017), 4051–4060.

Cronan, T. P., Al-Rafee, S., 2008. Factors that Influence the Intention to Pirate Software and Media. *Journal of Business Ethics*, 78, pp. 527–545.

Devine, D., Gaskell, J., Jennings, W. and Stoker, G., 2020. Exploring trust, mistrust and distrust (Unpublished work). University of Southampton, UK.

Dillard, J. P., 1994. Rethinking the Study of Fear Appeals: An Emotional Perspective. *Communication Theory*, 4(4), pp. 295-323.

Durojaiye, T., Mersinas, K. and Watling, D., 2020. What Influences People's View of Cyber Security Culture in Higher Education Institutions? An Empirical Study. The Sixth International Conference on Cyber-Technologies and Cyber-Systems.

Emery S. L., Szczypka, G., Abril, E. P., Kim, Y., Vera, L. 2014. Are you scared yet? Evaluating fear appeal messages in tweets about the Tips Campaign. *Journal of Communication*, 64, pp. 278-295.

Eyal, N., 2014. Hooked: How to build habit-forming products. Penguin.

Fishbein, M., Ajzen, I., 1975. Belief, attitude, intention, and behavior: *An introduction to theory and research*. Reading, MA.

Florêncio D. and Herley C. and Shostack A. 2014. FUD: A plea for intolerance. *Communications of the ACM*, 57(6), pp. 31-33.

Floyd, D.L., Prentice-Dunn, S. and Rogers, R.W., 2000. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology*, 30(2), pp.407-429.

Fogg, B. J., 2009, April. A behavior model for persuasive design. In Proceedings of the *4th International Conference on Persuasive Technology ACM.*

Gigerenzer, G., Todd, P. M., and the ABC Research Group. 1999. *Simple heuristics that make us smart*. Oxford.

Godin, G. and Kok, G., 1996. The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion*, 11, 87-98.

Goel, S., Williams, K., Huang, J., & Warkentin, M. 2020. Understanding the Role of Incentives in Security Behavior. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 3, 4241–4246.

Grüne-Yanoff, T., & Hertwig, R. 2016. Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26, 149–183.

Guitart-Masip, M., Duzel, E., Dolan, R. and Dayan, P., 2014. Action versus valence in decision making. *Trends in cognitive sciences*, 18(4), pp.194-202.

Hale, J.L., Householder, B.J. and Greene, K.L., 2002. The theory of reasoned action. *The persuasion handbook: Developments in theory and practice*, *14*(2002), pp.259-286.

Hansen, P. G. and Jespersen, A. M., 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behavior change in public policy. *European Journal of Risk Regulation*, 4(1), pp. 3-28.

Hastings, G., Stead, M., Webb, J., 2004. Fear appeals in social marketing: Strategic and ethical reasons for concern. *Psychology and marketing*, 21 (11), pp. 961-986.

Hausman, D. M., Welch, B., 2010. Debate: To nudge or not to nudge. *Journal of Political Philosophy*, pp. 123-126.

Held, V., 2006. The Ethics of Care: Personal, Political, Global. Oxford.

Hertwig, R. and Grune-Yanoff, T., 2017. Nudging and boosting: steering or empowering good decisions. *Perspectives on Psychological Science*, 12, pp. 973-986.

Hogarth, R. M., Soyer, E., 2015. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition*, 4, 221–228.

Humaidi, N. and Balakrishnan, V., 2015. Leadership styles and information security compliance behavior: The mediator effect of information security awareness. *International Journal of Information and Education Technology*, 5(4), p.311.

Hursthouse, R., 1996. “Normative Virtue Ethics.,” In R. Crisp (ed.): How should one live?. Oxford, pp. 19-36.

Hyman, M. R., and Tansey, R., 1990. The Ethics of Psychoactive Ads. Journal of Business Ethics 9(2), pp. 105-114.

Johnston, A. C., Warkentin, M., 2010. Fear Appeals and Information Security Behaviors: An Empirical Study. *MIS Quarterly*, 34(3), pp. 549-566.

Johnston, A. C., Warkentin, M., & Siponen, M. 2015. An enhanced fear appeal rhetorical framework: Leveraging threats to the human asset through sanctioning rhetoric. *MIS Quarterly*: *Management Information Systems*, 39(1), 113–134.

Kahneman, D., 2011. *Thinking fast and slow*. New York.

Kant, I., 1998 [1785]. *Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Kraemer, S., Carayon, P., Clem, J., 2009. Human and Organizational Factors in Computer and Information Security: Pathways to Vulnerabilities. *Computers & Security*, 28, pp. 509-520.

Leventhal, H., 1970. Findings and theory in the study of fear communications. In L. Berkowitz (ed.). *Advances in experimental social psychology*. Vol. 5. New York, pp. 119-186.

Martin, R. A., 2010. The Psyhology of Humor: An Integrative Approach. Burlington MA.

McGuire, W., 1968. Personality and Attitude Change: An Information Processing Theory. In A. G. Greenwald, T. C. Brock and T. M. Ostrom (eds.). *Psychological Foundations of Attitudes*, pp. 171-196.

Mersinas, K., Sobb, T., Sample, C., Bakdash, J.Z. and Ormrod, D., 2019. Training Data and Rationality. *Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics* (p. 225).

Mersinas, K. and Chana, C.D., 2022. Reducing the Cyber-Attack Surface in the Maritime Sector via Individual Behaviour Change. The Seventh International Conference on Cyber-Technologies and Cyber-Systems.

Mill, J. S. 1859. Utilitarianism. London.

Milne, S., Sheeran, P. and Orbell, S., 2000. Prediction and Intervention in Health-related Behavior: A Meta-analytic Review of Protection Motivation Theory. *Journal of Applied Social Psychology,* 30(1), pp. 106-143.

Münscher, R., Vetter, M., Scheuerle, T., 2016. A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, 29(5), pp. 511-524.

Nickerson, C., 2022. *Theory of Reasoned Action.* Available at: https://www.simplypsychology.org/theory-of-reasoned-action.html (Accessed: 22/12/2022).

Nisbett, R. 2004. The geography of thought: How Asians and Westerners think differently - and why. London.

Nussbaum, M., 2001. Upheavals of Thought: The Intelligence of Emotions. Cambridge University Press.

Parkinson, J., & Haggard, P., 2014. Subliminal priming of intentional inhibition. *Cognition*, 130(2), 255–265.

Pinder, C., Vermeulen, J., Cowan, B. R. and Beale, R., 2018. Digital Behavior Change Interventions to Break and Form Habits. *ACM Transactions on Computer-Human Interaction*, 25(3), 66 pages.

Reid, R., van Niekerk, J., 2016. Decoding Audience Interpretations of Awareness Campaign Messages. *Information and Security*, 24(2), pp. 177-193.

Renaud, K., Dupuis, M., 2019. Cyber Security fear appeals: unexpectedly complicated. *New Security Paradigms Workshop* (NSPW '19), September 23-26 2019.

Rogers, R. W., 1975. A protection motivation theory of fear appeals and attitude change. *Journal of Psychology*, 91, pp. 93-114.

Rogers, R. W., 1983. Cognitive and Psychological Processes in Fear Appeals and Attitude Change: A Revised Theory of Protection Motivation. Social Psychophysiology: A Sourcebook, pp. 153-176.

Ruiter, R. A. C., Kessels, L. T. E., Peters, G-J., Kok, G., 2014. Sixty Years of Fear Appeal Research: Current State of the Evidence. *International Journal of Psychology*, 49(2), pp. 63-70.

Saghai, Y., 2014. Salvaging the concept of nudge. *Journal of Medical Ethics*, 38, pp. 487-493.

Shabel, S. J., Wang, C., Monk, B., Aronson, S., Malinow, R., 2019. Stress Transforms Lateral Habenula Reward Responses Into Punishment Signals. *Proceedings of the National Academy of Sciences of the United States of America*, 116(25), pp. 12488-12493.

Simon, H. A., 1972. Theories of bounded rationality. In C. B. McGuire and R. Radner (eds.). *Decision and Organization,* pp. 161-176.

Simon, H. A., 1991. Bounded rationality and organizational learning. *Organization Science*, 2(1), pp. 125-134.

Siponen, M., Iivari, J., 2006. Six Design Theories for IS Security Policies and Guidelines. *Journal of the Association for Information Systems*, 7(7), pp. 445-472.

Siponen, M., and Vance, A. O., 2010. Neutralization: New Insights into the Problem of Employee Systems Security Policy Violations. MIS Quarterly, (34: 3), pp. 487-502.

Siponen, M., Mahmood, M. A. and Pahnila, S., 2014. Employee's Adherence to Information Security Policies: An Exploratory Field Study. *Information and Management,* 51, pp. 217-224.

Staats, H. in Spielberger, C., 2004. Encyclopedia of applied psychology. Academic press.

Suh, M. M. and Hsieh, G., 2016. Designing for Future Behaviors: Understanding the Effect of Temporal Distance on Planned Behaviors. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1084-1096.

Sunstein, C. R. 2014. Why nudge? The politics of libertarian paternalism. New Haven CT.

Sunstein, C. R. 2016. The ethics of influence. Government in the age of behavioral Science. New York.

Sutton, S. R., 1982. Fear-arousing communications: A critical examination of theory and research. In J. R. Eiser (ed.). *Social psychology and behavioral medicine*. London, pp. 303-337.

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracin, D., 2015. Appealing to fear: A Meta-Analysis of fear appeal effectiveness and theories. *Psychological Bulletin, 141,* 1178–1204. http://dx.doi.org/10.1037/a0039729

Tengland, P. A., 2012. Behavior change or empowerment: on the ethics of health-promotion strategies. *Public Health Ethics*, 5 (2), pp. 140-153.

Thaler, R. H. and Sunstein, C. R., 2008. Nudge. Improving Decisions about Health, Wealth and Happiness. Yale.

Theocharidou, M., Kokolakis, S., Karyda, M., Kiountouzis, E. 2005. The insider threat to information systems and the effectiveness of ISO17799. *Computers and Security*, 24, pp. 472-484.

Thomson, K., and van Niekerk, J., 2012. Combating information security apathy by encouraging prosocial organizational behavior. *Information Management and Computer Security*, 20(1), 39–46.

van Bavel, R., Rodríguez-Priego, N., Vila, J., Briggs, P., 2019. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human-Computer Studies*, 123, pp. 29-39.

van den Akker, L., Heres, L., Lasthuizen, K., Six, F., 2009. Ethical Leadership and Trust: It's All about Meeting Expectations. *International Journal of Leadership Studies*, 5(2), pp. 102-122.

Van Staveren, I., 2007. Beyond utilitarianism and deontology: Ethics in economics. *Review of Political Economy*, *19*(1), pp.21-35.

Waldrop, M. M., 2016. How to hack the hackers: the human side of cybercrime. Nature, 533 (7602).

Wallston, K. in Smelser, N.J. and Baltes, P.B. eds., 2001. International encyclopedia of the social & behavioral sciences (Vol. 11). Amsterdam: Elsevier.

Webb, T. L. and Sheeran, P., 2006. Does Changing Behavioral Intentions Engender Behavior Change? A Meta-analysis of the Experimental Evidence. *Psychological Bulletin,* 132(2), pp. 249-268.

Wegner, D. 2002. The illusion of conscious will. London.

Weinstein, N. D. (1993). Testing four competing theories of health-protective
behavior. *Health Psychology,* **12,** 324-333.

Weirich, D., Sasse, M. A., 2001. Persuasive Password Security. CHI, 139-140.

Wilkinson, T. M. 2013. Nudging and manipulation. *Political Studies,* 61, 341–355.

White, M., 2013. The manipulation of choice: Ethics and libertarian paternalism. Springer.

Witte, K., 1992. Putting the Fear Back Into Fear Appeals: The Extended Parallel Process Model. *Communications Monographs* 59(4), pp. 329-349.

Witte, K. and Allen, M., 2000. A meta-analysis of fear appeals: implications for effective public health campaigns. *Health Educ. Behav.*, 27 (5), pp. 591-615.

Wood, W., Quinn, J.M. and Kashy, D.A., 2002. Habits in everyday life: thought, emotion, and action. *Journal of personality and social psychology*, 83(6), p.1281.

Workman, M., Bommer, W. H. and Straub, D., 2008. Security Lapses and the Omission of Information Security Measures: A Threat Control Model and Empirical Test. *Computers in Human Behavior,* 24, pp. 2799-2816.

Yeung, K., 2012. Nudge as Fudge. The Modern Law Review, 75 (1), pp. 122-148.