Entrapment behind the firewall: the ethics of internal cyber-stings

Morgan Luck
Charles Sturt University
moluck@csu.edu.au

Abstract

Internal cyber-attacks (cyber-attacks which occur from within an organization) pose a serious threat to an organization's security. One tool that organizations can employ to help them detect such threats is the internal cyber-sting. An internal cyber-sting involves an organization enticing its members into performing a (controlled) internal cyber-attack in order to apprehend them. However, there is (rightly) considerable moral consternation about employing such a tool; for it is deceitful and undermines trust. The aim of this paper is to present four separate actions that might be taken by organizations to strengthen their moral reason for employing internal cyber-stings.

Keywords: Entrapment; Cyber-Security; Stings; Internal Threats; Ethics

1 Introduction

Internal cyber-attacks (cyber-attacks against an organization from within the organization) pose a serious threat to an organization's security. IBM's *Cyber Security Intelligence Index* (2016) claims that 60% of all cyber-attacks are internal. Verizon's *Data Breach Investigations Report* (2016) claims that 77% of cyber-attacks that result in a breach involve an insider (p.36). These attacks are not only frequent, but also very damaging; with the estimated cost to organizations for each successful internal attack being between $100,000 and $500,000 (USD) (CA Tech., p.15, 2018). In response, organizations are justifiably "shifting their focus on detection of insider threats" (p.4). One tool that organizations can employ to help them detect these internal threats is an *internal cyber-sting*. An internal cyber-sting involves an organization enticing its members into performing a (controlled) internal cyber-attack in order to apprehend them. However, there is (rightly) considerable moral consternation about employing such a tool; for (amongst other concerns) it is deceitful and undermines trust. The aim of this paper is to present four separate actions that might be taken by organizations to strengthen (but perhaps not justify) their moral reason for employing internal cyber-stings.

To this end, this paper is broken into several sections. In the next section, section 2, I describe internal cyber-stings. In section 3, I relate these stings to entrapment (arguing that they are instances of neutral entrapment). In section 4, I explain what is meant by strengthening an organization's moral reason for the use of stings. In the subsequent four sections, 5 to 8, I will introduce each of the four actions in turn, and the principles which motivate them. We proceed now to a description of internal cyber-stings.

2 What are internal cyber-stings?

In order to introduce internal cyber-stings, consider the following paradigm case:

> A banking security manager decides to employ a hacker to help them expose weaknesses in their online banking system. However, the manager has a concern. The

> hacker's work requires that she has access to important internal systems. The manager worries that the hacker, given her abilities, could exploit this access and secretly transfer large sums of money to untraceable accounts. In order to help protect the bank's interests the manager creates a financial account in a vulnerable location that the hacker will surely notice. However, unbeknownst to the hacker, if she attempts to improperly transfer funds out of this account, security will be instantly informed and they will be detained. The hacker then proceeds to transfer the funds out of the account.

This *hacking case* has at least three salient features. First, it involves a type of sting – that is, an agent (in this case the manager) is enacting a plan that uses deceit to ensnare a target (the hacker). Second, the sting is internal – that is, the target and the agent are both part of the same organization (the bank), and the sting is performed in its interests. Third, the sting largely unfolds in a cyber-environment (the bank's IT systems). These are the three main features of internal cyber-stings.

These types of stings are strongly related to entrapment. Consequently, we can build upon the reasons given for the moral impermissibility of entrapment to help us judge the moral status of internal cyber-stings. To this end, we shall proceed to relate this type of sting to entrapment.

3 How do internal cyber-stings relate to entrapment?

There are important similarities between internal cyber-stings and entrapment which, if made clear, should help us to consider the moral permissibility of these stings. In this section, I introduce some of these similarities (and potential differences).

Entrapment is most commonly understood as a legal defence to criminal charges. It is used when, roughly speaking, an officer of the law (the agent) performs a sting upon a target, which induces them to perform some (otherwise criminal) wrongdoing that they would not have been likely to perform if not for the sting. If this defence is successful (that is, if the sting is determined to be an instance of entrapment), the target is found to be *not guilty* of the criminal charge. I shall refer to this understanding of entrapment as *legal entrapment*.

Legal entrapment is understood negatively. That is, it is something which law officers should not do. Not simply because, if proven, no conviction can be forthcoming; but because it is morally wrong. Reasons for this vary. One reason is that not only do such stings fail to prevent crimes (as it is unlikely the target would have committed this type of crime had they not been entrapped). A second reason is that they cause people to act wrongly, and as result makes the world worse. A third reason is that legal entrapment "violates the autonomy of those subject to it, it undermines an essential condition of moral agency and criminal liability" (Hughes, 2007, p.45). A forth reason is that "such activity will only lead to public disillusionment with our police and criminal justice system and a decrease in respect for the law" (Carlon, 2007, p.1101). A fifth reason is "because it subverts the moral capacities of entrapped persons" (Howard, 2016, p.25). These reasons are far from exhaustive. However, not all conceptions of entrapment are negative.

Hill, McLeod and Tanyi (2017) promote a neutral, and broader, understanding of entrapment. It is neutral in the sense of being morally neutral (i.e. stings are not necessarily morally permissible, nor are they necessarily impermissible). And it is broader in that (a) the agents

involved need not be law officers (that is, it includes instances of private entrapment),[1] and (b) the actions of the target need not be otherwise criminal. Hill et al. argue that the entrapment occurs whenever:

> 1. an agent plans that the target commit an act;
> 2. the planned act is of a type that is criminal, immoral, embarrassing, or socially frowned upon (measurable in part by the extent to which the target would probably not like the act to be exposed to colleagues, an employer, friends, family, or the public);
> 3. the agent procures the act (by solicitation, persuasion, or incitement);
> 4. the agent intends that the target's act should, in principle, be traceable to the target either by being detectable (by a party other than the target) or via testimony (including the target's confession), that is, by evidence that would link the target to the act;
> 5. in procuring the act, the agent intends to be enabled, or intends that a third party should be enabled, to prosecute or to expose the target for having committed the act.
>
> (pp.13 – 14)

To illustrate this account, let us see how it applies to the hacking case. First, the bank's security manager (the agent) plans that the hacker (the target) commit an act (that of transferring the funds). Second, the transfer of the funds is the type of act that is clearly immoral/criminal. Third, the manager procures the transfer (by placing the vulnerable account in a place the hacker would surely notice). Forth, the manager intends that the hacker's transfer be traceable to the hacker. Fifth, in procuring the transfer, the manager intends to expose the hacker. So, according to Hill *et al*. this is a case of entrapment, in the neutral, broad sense. I shall refer to this understanding of entrapment as *neutral entrapment*.

Although the paradigm case of an internal cyber-sting (the hacking case) is also an instance of neutral entrapment, it does not follow that all cases of internal cyber-stings are. It may be that the four actions I will go on to establish (for strengthening an organization's moral reason to undertake a sting) are applicable to all cases of internal cyber-stings; yet, for the sake of simplicity we focus solely on internal cyber-stings that are also instances of neutral entrapment.

So, with this sub-class of internal cyber-stings identified, and their relation to entrapment made clear, we are now well placed to consider what actions might strengthen an organization's moral reason to enact such stings. However, before outlining these actions, let's first be clear about what is meant by "strengthening moral reason".

4 What does it mean to strengthen moral reason?

The aim of this paper is to detail four actions which organizations might take to strengthen their moral reason for employing internal cyber-stings. But what exactly is a moral reason and what does it mean to strengthen one's moral reason (or to have stronger moral reason)? A

[1] See Yaffe (2005) for more on the notion private entrapment.

moral reason is a reason for the moral permissibility of an action (a reason which has some degree of moral weight). For example, a moral reason for a surgeon to cut their patient is that it is required remove a malignant tumour. Moral reason for an action is strengthened when we have more reason for its moral permissibility. For example, a surgeon's moral reason for cutting a patient is strengthened when, in addition to needing to cut out the malignant tumour, they also have the patient's consent to being cut. Although this may sound straightforward, it is worth noting that: (1) strengthening moral reason for an action doesn't necessarily make it morally permissible; (2) it is possible to strengthen the moral reason for an action that is already permissible; and (3) that the notion of stronger moral reason can operate under multiple ethical theories. The reminder of this section will focus on these three points.

The first point to note is that, strengthening the moral reason for an action doesn't necessarily make it morally permissible. To illustrate this point, consider again the surgeon who discovers that their patient has a malignant tumour that needs to be cut out. This discovery strengths the surgeon's moral reason to cut their patient. However, this doesn't mean its permissible. Why? Because, the surgeon also requires the consent of the patient for the procedure to be permissibly. So, there is a marked difference between having more reason for an act's permissibility, and the act *being* permissible; or as Norcross (2006) puts it, the "fact that there is a moral reason to perform some action, even that there is more moral reason to perform it than any other action, doesn't mean that one ought to perform it" (p.48). This point is worth stressing, for the arguments presented here are *not* for the moral permissibility of internal cyber-stings; they merely aim to provide more moral reason for their permissibility.

The second point worth noting is that, it is possible to strengthen the moral reason for an action that is already permissible. To illustrate this point, consider again the surgeon who discovers that their patient has a malignant tumour that needs to be cut out. Let us assume the surgeon has fulfilled all the conditions required for the operation to be morally permissible (e.g. gained the patient's consent, etc.). However, it may be still be possible to strengthen the surgeon's moral reason for the operation. For example, imagine that the surgeon is informed that the patient is a scientist who is on the brink of curing cancer; and without the operation they will be unable to complete the cure. If this is an additional reason for the permissibility of removing the tumour, then the surgeon's moral reason has been strengthened. However, removing the tumour would have been permissible regardless of this stronger moral reason. This point is worth stressing, as although the arguments presented here aim to strengthen the moral reason for internal cyber-stings, there is no presumption that such stings are impermissible without such strengthening.

The third and last point worth noting is that the notion of stronger moral reason can operate under multiple ethical theories. To illustrate this let us consider an example of stronger moral reason operating under consequentialism, and another under deontological ethics.

Consequentialism is the ethical theory "that whether an act is morally right depends only on the consequences of that act" (Sinnott-Armstrong, 2015). Broadly put, consequentialism holds that the act that results in the most moral value is the act you should perform. To illustrate how one's moral reason might be strengthened under consequentialism, consider the following case,

> A surgeon in an understaffed ER ward who must choose between operating on patient A or patient B (where the patient left untreated will die).

For the sake of simplicity let us proceed on the assumption that the lives at stake in this example are the only things of moral value, and every life is equally valued. As things stand, saving either patient seems morally permissible to the surgeon, as in either case she saves the most lives she can (i.e. one life). However, upon hearing that patient A is a scientist who is on the brink of curing cancer, their moral reason for saving patient A has now been strengthened. This is because, more lives will be saved if patient A survives. This an example of moral reason being strengthened under consequentialism. Let us now consider an example of stronger moral reason under deontological ethics.

According to deontological ethics, "what makes a choice right is its conformity with a moral norm" (Alexander, 2016). A moral norm can be thought of as a rule that must be followed (regardless of the consequences) to act morally. To illustrate how one's moral reason might be strengthened under deontological ethics consider following example,

> You have the option to either do nothing and let an empty runaway trolley careen down track A killing three innocent people stuck there, or instead redirect the trolley to track B where it will safely run off the tracks, over a cliff and into the sea, or instead redirect the trolley to track C where it will safely come to a halt upon the track.

Let us imagine that the following two moral norms are in place: "Save lives" and "Do not be wasteful". And for the sake of simplicity, let us assume that these are the only relevant moral norms. In this case one might consider that there is stronger moral reason for redirecting the train to track C, rather than track B. Why? Because by redirecting the train to track C one upholds both moral norms (i.e. three lives are saved, and the trolley is not wasted). This an example of stronger moral reason under deontological ethics.

That moral reason can be strengthened (or can be stronger) under multiple ethics theories is worth stressing. For the arguments presented (which aim to strengthen the moral reason for internal cyber-stings) do no assume the correctness a particular ethical theory, or that they will operate equally well under multiple ethical theories.

So, given this understanding of what is means to have stronger moral reasons to perform an action, let us now proceed to outline each of the four actions; each of which I argue strengthens an organization's moral reason to employ an internal cyber-sting.

5 Action 1: Obtain the target's consent

In this section, I shall argue that there is stronger moral reason for an internal cyber-sting if the target properly consents to it. So, gaining the target's consent is an action an organization can perform to strength its moral reason for stinging. To this end, I will draw an analogy between internal stings and internal monitoring.[2]

Consider the following example of internal monitoring:

> A manager wishes to determine how hard her employees are working. Without notifying her employees of the possibility of this type of monitoring, she starts secretly reading their emails in order to determine to what extent they are business-related.

[2] Although not all internal monitoring occurs in the workplace (for example monitoring one's own family) since the most frequently considered type of internal monitoring is workplace monitoring, I will focus chiefly on this sub-class.

In this email-monitoring case, the manager's covert monitoring of her employees is generally considered to be impermissible. Reasons for this vary. Some (Miller & Weckert, 2000) argue that such a practice constitutes a breach of an individual's right to privacy (and in turn precipitates the breakdown of trust). Some (Clarke, 2005) argue that this practice fails to uphold the principle of autonomy. These reasons are likely to collapse into each other if, as Wall (2011) maintains, the moral right to privacy is built upon the moral right to autonomy (p.69). We will proceed on this assumption.

If covert internal monitoring is impermissible because it ultimately represents a breach of privacy (and so a breach of autonomy), a key question to consider is: would the manager's actions be less objectionable were she to gain her employees' consent to being monitored? Let us consider some reasons for why it would be less objectional.

First, if covert internal monitoring is impermissible because it constitutes a breach of an individual's right to privacy, then provided an employee provides proper consent (i.e. they are not improperly coerced, and they are sufficiently informed, etc.)[3] the employer may be able to avoid breaching this right. Why think this? Because, as McCloskey (1980) states,

> *It is evident that we can consent to forgo privacy, properly, as in marriage, friendship, and certain other social relationships, putting certain areas of our lives outside the sphere of privacy for certain other persons.* (p.37)

To illustrate this point, consider the right to privacy I have in my own home. I can consent to forgo some degree of privacy by choosing to live-stream my home-life antics on the web (provided I have not been coerced into this, and I am fully informed of the intrusion [i.e. where the cameras are etc.]). So how is it that by gaining someone's consent, we avoid breaching their right to privacy? Westin & Ruebhausen's seminal work on privacy provides an indication.

Westin & Ruebhausen define privacy as "the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others" (Westin & Ruebhausen, p. 7). And how do agents govern this extent. Wall argues that is via consent, casting privacy as the "moral right to consent to access by others to one's personal information" (p.69). So, by gaining someone's consent to monitor them, we uphold their moral right to consent, and so in turn uphold their right to privacy, and so avoid the breach.[4] Note that I am not suggesting that gaining proper consent is sufficient for internal monitoring to be permissible (although it may well be). Rather I am suggesting that gaining proper consent (as opposed to not) provides stronger moral reason for workplace monitoring.

One might think that this is obviously the case, on the grounds that if someone consents to some manner of treatment then it is permissible to treat them in this manner. However, this principle seems too strong, for, as Miller (2005) points out "even if people do consent to some

---

[3] See Palm (2009) for a discussion on the necessary conditions for morally acceptable consent (in the context of limiting one's right to privacy in the workplace).

[4] We can run a similar line if covert internal monitoring is impermissible because it fails uphold the moral right to autonomy (but not because of issues of privacy). For provided an employee properly consents, then one's ability "to determine the course of one's own life free of the interference of others" (Clarke, p.237), is diminished less by internal monitoring if one has freely chosen a course of life involving such monitoring.

sort of treatment, it doesn't follow that it is moral to treat them in that manner" (p.276). For some manners of treating people, for example treating people as slaves, may be impermissible regardless of the fact they have properly consented to such treatment. Yet what seems more defensible is the following principle: there is stronger moral reason for treating someone in a manner they have properly consented to, than otherwise. For example, although treating someone as a slave may *always* be morally impermissible, there is more moral reason to treat them as such if they have consented to it, than had they not. To support this point, imagine you were forced between treating someone as a slave who has properly consented to this (i.e. not coerced economically or otherwise [a submissive,[5] for example]), or treating someone as a slave who has not. This does not seem like a choice one should make by flipping a coin. The addition of consent does seem to have some moral weight.[6]

If this principle holds, then we can clearly see how it also applies to workplace stings. For, if the target properly consents to the possibility of a sting (as opposed to not), there will be more moral reason for its permissibility. So, in the hacking case, if the hacker properly consented to the possibility of a sting (perhaps as a standard condition of employment), there will be more reason for the sting's permissibility. However, even if this principle fails an argument can be given for the same conclusion by drawing an analogy between internal monitoring and internal stings.

If internal monitoring without consent is wrong because it breaches an employee's moral right to autonomy, then the question to consider now is why an internal sting without consent is wrong. To answer this question it is worth noting, as Dworkin (1988) does in relation to the stings used in cases of legal entrapment, that what "these techniques have in common is the use of deception to produce the performance of a criminal act under circumstances in which it can be observed by law enforcement officials" (p.130). That these stings involve deceit (by which the agent procures the wrongdoing) is the relevant point here; as deceit, it is sometimes argued, threatens autonomy. Why? One reason is that a deceptive environment is an environment where it is difficult make informed choices. And one needs to be able to make such choices in order to properly determine the course of one's own life. So, internal stings may be wrong because they involve deception, which threatens our autonomy.

If both internal stings without consent, and internal monitoring without consent, are wrong because they both breach the employee's autonomy, then perhaps, like internal monitoring, there can be more moral reason for internal stings when employers gain proper consent. In other words, perhaps the following argument can be mounted:

1. Internal monitoring and internal stings are relevantly similar.
2. There is stronger moral reason for internal monitoring when proper consent is gained (as opposed to when it isn't)

[5] A person who desires (at least for limited period of time) to submit their will to another.

[6] Note I am not suggesting that slavery might be permissible if consent is given (slavery may always be impermissible), only that consent has some moral weight. I am also not suggesting the permissibility of slavery is relative to the people being enslaved. As slavery may always be impermissible regardless of whether consent is given (despite it having some moral weight).

So,

3. There is stronger moral reason for internal stings when proper consent is gained (as opposed to when it isn't)

However, if this argument also fails to persuade, there is a further argument for the same conclusion we shall also now consider.

Rather than drawing an analogy between monitoring and stings, perhaps it might be more fruitful to think about how gaining proper consent reduces the element of deceit. To this end, let us consider a related issue in medical ethics: the use of placebos.

A placebo is "a substance that the physician believes has no known specific pharmacological activity against the condition being treated" (Bostick et all., 2008, p.58). Physicians sometimes administer placebos to patients without the patient knowing a particular treatment is a placebo, as this can have therapeutic benefits (known as the placebo effect). Some argue that it is impermissible to administer placebo treatments as they involve deception, and so "essentially run counter to the ideal of informed consent, which is central to the ideal of patient autonomy" (Groll, 2011, p.198). However, others argue that in those cases where it is possible to gain a patient's *general* consent, the treatment may be permissible,

> *For example, a physician could explain to a patient that a more certain diagnosis or better understanding of his or her condition could be achieved by evaluating the effects of different types of medication, including one that is not pharmacologically active, namely, a placebo. By obtaining the patient's cooperation in this manner, the physician need neither identify which medication is the placebo nor seek specific consent immediately before its administration. This example of shared decision making demonstrates an approach that respects the autonomy of patients and fosters trust within the patient-physician relationship. Moreover, the authorized use of placebos is not expected to significantly diminish their clinical effectiveness as research suggests that little variation in clinical outcomes is observed between patients who are informed that they are to be treated with placebos and patients who are administered placebos in a deceptive manner*. (Bostick et all., p.59)

The relevant point here is that the wrongness of a deceptive practice (a practice in tension with the patient's autonomy) is arguably lessened (or perhaps even alleviated) when proper consent to the practice is obtained. By the same token, there is reason to think that the wrongness of the deceptive practice involved in an internal cyber-sting is lessened when proper consent to the practice is obtained in a relevantly similar manner. Which in turn supports the conclusion that there is stronger moral reason for an internal cyber-sting if the target properly consents to it.

So, getting an employee to consent to the possibility of a sting is an action that organizations can take to strengthen the moral reason for their use. (The other additional benefit worth briefly mentioning is that informing employees of possible stings may also act as a deterrent.) However, some organizations may not wish to gain consent on the grounds that informing a target of the prospect of a sting may increase the possibility of them evading it. This is an understandable concern. However, even if the possibility of stinging a target is reduced in this fashion, it might not be reduced so much as to make stings ineffective. For example, although criminals outside of the workplace are aware of the possibility of a sting by law enforcement agencies, it doesn't follow that such stings fail to ensnare their targets. So, although gaining

consent may lower the chances of a successful sting, it may still be worth it given it strengths an organization's moral reason for the practice.

6 Action 2: Ensure the stakes are high

In this section, I shall argue that the more that is morally at stake, the stronger moral reason for the internal cyber-sting. So, ensuring that the stakes are high is something an organization can do to (proportionally) strengthen its moral reason for stinging. To help make this point, we begin with an illustration of a simpler principle.

Consider the following case:

> While discussing beloved films with a new acquaintance you are asked about your favourite film. You are about to tell her that it is *Bladerunner* when you remember that this is the password to your World of Warcraft account (a popular online game that you are mildly amused by). In order to lessen the chances of this acquaintance gaining access to your game account, you lie and tell her your favourite movie is *Akira* instead.

Some may argue that lying in this *password case* represents a breach of trust and an act of deceit, and so is to some extent (that is, pro tanto) morally wrong. Whether the breach might also be all-things-considered morally wrong, and so morally impermissible, is less clear. Perhaps, because what was at stake was so trivial it was impermissible to lie; or perhaps because what you were lying about was so trivial it was permissible. However, what does seem clear is that there is more reason to lie if "Bladerunner" were the password to your bank account (filled with your life's savings). The principle I am supporting here is: the greater the cost, the more reason there is to lower the risk.[7]

If this principle holds, then the more you stand to lose from your new acquaintance gaining access to your bank account, the more *reason* there is to lie. However, it does not follow that you have more *moral* reason to lie (or that it is morally permissible to lie). Whether there is more moral reason will depend on whether what is at stake is of moral import. In this password case, if there isn't anything morally at stake then there may be no moral reason to lie to your friend. But if there were something morally at stake then there would be moral reason to lie. This idea can be captured by the following principle (which is a variant of the previous simpler one): the greater the moral cost, the more moral reason there is to lower the risk.

If this principle holds then there will be stronger moral reason for a sting the more there is morally at stake (given the sting is a means to lower the risk). So, for an organization's moral reason to be strengthened by this consideration significantly, they should ensure that there is something of significant moral value at stake. Sometimes, it will be easy to determine whether there is something of significant moral value at stake. For example, a business that merely wishes to lower the amount of toilet paper their employees are using at work does not have anything of significant moral value at stake; so, it has little moral reason (in terms of the stakes) to enact a sting to discover culprits using generous amounts to toilet paper. However, agencies that protect sensitive security data (such as the CIA, MI5 or ASIO) clearly do have something

[7] Where cost is simply understood as something bad, and risk is understood as the chance of something bad.

of significant moral value at stake; so, they have much more moral reason to enact a sting in order to discover culprits who might seek to divulge this data to hostile groups.

Let us now return to internal cyber-stings and examine how this principle might be applied to the hacking case. Before enacting a sting upon the hacker, the banking security manager should consider whether there is anything morally at stake. At first it might seem that there is little at stake here morally, as the bank only stands to lose money, which (unlike a person's life) doesn't obviously have moral value. However, this assessment is far too quick.

The bank might have a moral obligation to its employees and/or shareholders to protect its own interests; and so, since both its financial and reputational interests might be damaged by the hacker's cyber-theft, the bank might have a moral reason to guard against it. And it is easy to see that the more money that is at stake the more moral reason there is to protect it, which in turn strengthens the moral reason for a cyber-sting that offers such protection. The bank might also have a moral obligation to protect their customers' interests; for example, the private information kept by the bank on its customers (such as addresses, dates of birth, and transaction information) might also be things the bank has moral reason to protect (reason complementary to, but independent from, its own interests).

The bank might also have a moral obligation to act against cyber-theft regardless of what they stand to lose. Why think this? Because organizations may have a moral reason, as we all arguably do, to take reasonable steps to act against wrongdoing. Note that I am not suggesting that they have a moral obligation to stop wrongdoings (although presumably they sometimes do). Only that, they have some moral reason to. For example, if you pass someone about to wrongfully litter, then, even if you do not have a moral obligation to stop them, there still exists a moral reason for you to do so. And the bigger the wrongdoing the more moral reason there is for its prevention. Weckert and Miller make a similar point regarding the permissibility of internal monitoring, stating that "Theft and fraud in the workplace are still theft and fraud, so some surveillance can be justified in order to apprehend culprits" (p.260). Or put another way, if failure to prevent cyber-theft comes at a moral cost, there is some moral reason to prevent it.

Consequently, if this principle holds (that the greater the moral cost, the more moral reason there is to lower its risk) then the more that is morally at stake, the stronger the moral reason for taking steps to lower the risk; steps which may include enacting an internal cyber-sting. Note that I am not suggesting here that organizations should not enact an internal cyber-sting unless the stakes are high (although this may be the case). Only that by ensuring that the stakes are high, organizations can be ensured that there is stronger moral reason for a sting (as opposed to if the stakes are low).

7 Action 3: Ensure you have evidence of the target's predisposition to act wrongly

In this section, we shall argue that the more evidence an agent has for the target's predisposition to act wrongly (in a particular way), the more moral reason the agent has to sting the target (by activating this predisposition). So, ensuring that there is considerable evidence for the predisposition is something organizations can do to (proportionally) strengthen its moral reason for stinging.

To help establish this principle, consider the following case:

> You have no reason to think that your partner is cheating on you, however you are determined to put them to the test anyway. To this end you use your extensive knowledge of your partner to create an enticing online profile of their perfect match. You then contact your partner online under the guise of this perfect match in order to see if you can catfish[8] them.

Some may argue that your actions in this *catfish case* are morally impermissible. This is because (in addition to the use of deceit and the violation of trust) you are also attempting make your partner do something immoral (i.e. to intend to cheat on you).

Unless one is a moral saint, we are all vulnerable such traps. That is, even those of us who only have a small disposition (rather than a predisposition) to act wrongly, will act wrongly if this disposition is sufficiently encouraged. Given this, we shouldn't encourage people to act wrongly unless we have good reason; otherwise we are making the world worse. As Dworkin (1988) states in respect to the techniques lead to legal entrapment:

> *The central moral concern with pro-active law enforcement techniques is that they manufacture or create crime in order that offenders be prosecuted and punished. They do not discover criminal activity; they create it.* (p.136)

Or, as Carlon (2007) puts it, it is critical that "we are prosecuting existing crimes, rather than crimes that we have created for the purpose of punishment" (p.1124). So, there is a moral risk to encouraging wrongdoings via such stings, as there is a chance one is acting impermissibly by creating wrongdoings that would not have occurred otherwise.

Now let us contrast this catfishing case to one where we have some evidence that our partner has a predisposition to cheat (although this predisposition is not so strong that it mitigates any blame if acted upon; for example, he is not a sex addict). Let us also stipulate that, although we have some evidence that they cheat, we do not have sufficient evidence to justifiably end the relationship (and so save ourselves from being cheated on longer than is necessary). Although it still might prove impermissible to catfish our partner in this case, it does now seem less wrong. Or put another way, there seems to be more moral reason for the sting in this case. So why think this?

Although in both cases the sting causes the partner's disposition/predisposition to manifest, in the case of the partner with the small disposition to cheat, had they not been stung the chances of them intending to cheat would be low. Whereas in the case of the partner with the predisposition to cheat, had they not been stung, the chances of them intending to cheat would remain high. What does this tell us? It tells us it is likely that the first sting makes the world worse, whereas the second doesn't (in fact the second sting provides you with the evidence required to end the relationship and move on). This difference is moral one. That is, given the possibility of making the world worse, the more evidence we have that our target would have performed the wrongdoing in the absence of the sting, the more moral reason there is for the sting.

Accordingly, we can mount the following principle on the back of this difference: the more evidence an agent has for the target's predisposition to act wrongly (in a particular way), the more moral reason the agent has to sting the target (by activating this predisposition). This principle is also supported by one of the arguments for entrapment defence: the subjective

[8] That is, trick them into making a romantic advance with the fake online persona.

approach. The subjective approach holds that the entrapment defence is successful (resulting in the defendant being found not guilty) when the evidence fails to demonstrate that the defendant has a predisposition to act wrongly. And inversely, the "entrapment defence fails if it can be shown that the defendant was "predisposed" to commit the crime" (Yaffe, p.7).

Why? Because, as Sinnott-Armstrong (1999) states, those "with weak predispositions to crime are unwary innocents, whereas those with strong predispositions to crime are the very people who need to be found guilty and punished to prevent their crimes" (p.99). So, for example, although someone might fall for a sting by unwarily buying illegal drugs from an uncover police officer, if the evidence gleamed from this sting is insufficient to demonstrate a criminal predisposition to buy such drugs, they may be found not guilty, and the moral permissibility of a sting questioned.

On the other hand, if one has prior evidence (perhaps a criminal record, or testimonial evidence, of the defendant buying illegal drugs) (evidence that motivated the police stinging this individual), then if the sum of the evidence is sufficient to demonstrate the predisposition to buy the drugs, they may be found guilty, and the moral permissibility of a sting left unquestioned.

How does the subjective approach support this principle? The subjective approach holds that if one has sufficient evidence to demonstrate a target's predisposition to act wrongly then the sting is morally permissible. It follows from this that the more evidence one has of this predisposition, the more moral reason there is for the sting's permissibility.

This principle is of particular relevance to the hacking case (and perhaps the cybersecurity sector more generally). This is because there is a real pragmatic advantage to organizations employing hackers with actual cyber-crime experience (that is, current or former black and grey hat hackers) to help strengthen their security; for, as the adage attests, set a thief to catch a thief. However, with this pragmatic advantage also comes a risk, as the hacking case illustrates. But if one has evidence that the hacker once performed such wrongdoings, one also has, to some degree, evidence of a predisposition. Why is this important? Because if this principle holds, then the more evidence an organization has, the stronger the moral reason to target the hacker with a sting. Note that I am not suggesting that organizations should not enact an internal cyber-sting unless they have sufficient evidence of this predisposition (although this may be the case). Only that by ensuring that they have such evidence, organizations can be ensured that there is stronger moral reason for stinging.

8 Action 4: Ensure the sting's inducement is low

In this section, we shall argue that the lower the sting's inducement to act wrongly, the stronger the moral reason for the sting. So, ensuring that the inducement is low is something organizations can do to (proportionally) strengthen its moral reason for stinging.

To help establish this principle, consider the following case:

> Thomas Anderson, a struggling unemployed software developer, is attending a job interview at a prestigious tech company. Before facing the interview panel, the HR officer assigns Thomas to a computer to undertake a standard English language aptitude test. Thomas discovers that the previous computer's user is still signed into their email account. Although Thomas knows that the right thing to do is sign out of this account on their behalf, he is enticed by the subject heading of the most recent

> email: "Questions to ask Thomas Anderson during the interview today". Despite his better judgment Thomas opens and reads the email. What he doesn't realize is that he has been stung. The HR officer engineered this situation to expose (to the interview panel) each candidate's willingness to read someone else's emails.

Some might argue that the HR officer acted wrongly in the *interview* case. One reason for this is that the inducement to read the email was too high; that the chance for a struggling unemployed job-candidate to gain an advantage in an interview for a prestigious job is just too tempting. Let us grant that even if Thomas only had a very small disposition to violate someone's privacy, he would still be all but compelled to read this email.

So why is it that baiting a sting with such a strong inducement might be impermissible? It is for the same reasons as given in the previous section; that is, it is because, as Yaffe states (in regards to legal entrapment), these stings may be causing wrongdoings without preventing any.

> *…by engaging in behavior that risks ensnaring the unpredisposed…the government engages in behavior that is likely to cause criminal conduct without preventing any, since the unpredisposed are not likely to have acted criminally without temptation from the government.* (Yaffe, p16)

Let us take as read that causing such wrongdoings is wrong. We discussed (in the previous section) that one way to lower the chances of causing such a wrongdoing is by gathering evidence of the target's predisposition to act wrongly prior to establishing the sting. In this section, we focus on establishing a particular kind of sting with a low inducement, in order to lower the chances of causing the same type of wrongdoing (i.e. one that would not have otherwise occurred).

To see why it is that the lower the sting's inducement, the lower the chances of causing a wrongdoing that would not have otherwise occurred, we need to consider the relationship between predispositions and inducements. This relationship, as Hall *et al* explains, is one of inverse proportion:

> *Other things being equal, the greater the degree of inducement or enticement has to be before the person succumbs to temptation and commits the offence, the less the person can properly be held to have been predisposed to commit the offence.* (p.15)

Given this relation, as Carlon states, "the success of a slight inducement indicates great predisposition; the resort to an extremely powerful inducement militates for a finding of less predisposition" (p.1093). So, if we have more moral reason to sting those with higher predispositions to act wrongly, then we have moral reason to lower the sting's inducement (without rendering it impotent).

Let us return to the interview case to apply this principle. What this principle tells us is that, if the lure of subject heading "Questions to ask Thomas Anderson during the interview today" constitutes an inducement so high that Thomas was all but compelled to take a peek, then it offers little indication of him having a predisposition to breach someone's privacy. But it does more strongly indicate that the HR officer has done something wrong, by causing a wrongdoing where none would had occurred otherwise. Yet, were we to adjust the sting somewhat, so the subject heading was now something much less enticing, such as "How to get into the executive bathroom", and Thomas now took a peek, then this sting offers more of

an indication that he has a predisposition to breach someone's privacy, and in turn less strongly indicates that the HR officer has done something wrong.

So, let us apply this to the internal cyber-security hacking case. When setting up the sting, the bank's security manager has (according to this principle) moral reason to lower the amount of money in the account that the hacker might be tempted to steal. For the more money in the account, the more chance there is that, even if the hacker only had a small disposition to steal, he would be all but compelled to do so, and so the greater the chance the manager would have done something wrong in causing this wrongdoing.

Consequently, if the principle holds (that the lower the sting's inducement to act wrongly the stronger the moral reason for the sting) there is moral reason to setup the cyber-sting with a small (but still potent) inducement. Note that I am not suggesting by this that organizations should not enact an internal cyber-sting unless the inducement is small (perhaps if the stakes are high enough, for example, it might still be warranted). Only that by ensuring that the inducement is low, organizations can be ensured that there is stronger moral reason for the sting.

9 Conclusion

The aim of this paper was to present four actions (each underpinned by a separate principle) that can be taken by organizations to strengthen their moral reason for employing internal cyber-stings, and so increase their internal cybersecurity. These actions were: obtain the target's consent; ensure the stakes are high; ensure you have evidence of the target's predisposition to act wrongly; and ensure the sting's inducement is low. It is worth emphasising in closing that taking these actions is no guarantee that any subsequent sting is morally permissible (perhaps they never are). However, in the absence of an established account which clearly lays out the necessary and sufficient conditions for permissibility of neutral entrapment, it seems prudent to consider these recommended actions before employing an internal cyber-sting.

References

Alexander, L. (2016). "Deontological Ethics", In Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/ethics-deontological/

Bostick, N., Sade, R., Levine, M. & Stewart, D. (2008). "Placebo Use in Clinical Practice: Report of the American Medical Association Council on Ethical and Judicial Affairs", *The Journal of Clinical Ethics,* 1:19.

CA Tech. (2018) *Insider Threat Report*. CA Technologies.

Carlon, A. (2007). "Entrapment, Punishment, and the Sadistic State", *Virginia Law Review*, 93:1081.

Clarke, S. (2005). "Informed Consent and Electronic Monitoring in the Workplace", In Weckhert, J. (Ed.) *Electronic Monitoring in The Workplace: Controversies And Solutions,* Idea Group Publishing.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge University Press.

Groll, D. (2011). "What You Don't Know Can Help You: The Ethics of Placebo Treatment", *Journal of Applied Philosophy*, 2:28.

Hill, D.J., McLeod, S.K. & Tanyi, A. (2017). "The Concept of Entrapment", *Criminal Law and Philosophy* (2017). https://doi.org/10.1007/s11572-017-9436-7

Hughes, P.M. (2004). "What is wrong with entrapment" *The Southern Journal of Philosophy*, 1:42

IBM. (2016). *2016 Cyber Security Intelligence Index*. IBM x-force Research.

McCloskey, H. J. (1980). "Privacy and the Right to Privacy", *Philosophy* 55: 211.

Miller, S. (2005). "Guarding the Guards: The RIght to Privacy, and Workplace Surveillance and Monitoring in Policing", In Weckhert, J. (Ed.) *Electronic Monitoring in The Workplace: Controversies And Solutions,* Idea Group Publishing.

Miller, S. & Weckert, J. (2000). "Privacy, the Workplace and the Internet", *Journal of Business Ethics,* 3:*28*.

Norcross, A. (2006). "Reasons without demands: Rethinking rightness", In Dreier, J. (ed.), *Contemporary Debates in Moral Theory*. Blackwell.

Palm, E. (2009). "Securing privacy at work: The importance of contextualized consent", *Ethics and Information Technology,* 4:11.

Sinnott-Armstrong, S. (1999). "Entrapment in the Net?", *Ethics and Information Technology*, 2:1.

Sinnott-Armstrong, S. (2015). "Consequentialism", In Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/consequentialism/

Verizon. (2016). *Data Breach Investigations Report*. Verizon Enterprises.

Wall, E. (2011). "Privacy and the Moral Right to Personal Autonomy", *International Journal of Applied Philosophy* 1:25.

Westin, A. F., & Ruebhausen, O. M. (1967). *Privacy and freedom*. New York: Atheneum.

Yaffe, G. (2005). "'The Government Beguiled Me': The Entrapment Defense and the Problem of Private Entrapment", *Journal of Ethics and Social Philosophy*, 5:2.