

Sludge for Good: Slowing and Imposing Costs on Cyber Attackers

Josiah Dykstra
Trail of Bits
josiah.dykstra@trailofbits.com

Kelly Shortridge
Fastly, Inc.
kelly@shortridge.io

Jamie Met
National Security Agency
jlmet@nsa.gov

Douglas Hough
Johns Hopkins Bloomberg School of Public Health
douglas.hough@jhu.edu

Abstract

Choice architecture describes the design by which choices are presented to people. Nudges are an aspect intended to make “good” outcomes easy, such as using password meters to encourage strong passwords. Sludge, on the contrary, is friction that raises the transaction cost and is often seen as negative by users. Turning this concept around, we propose applying sludge for positive cybersecurity outcomes by using it offensively against attackers to consume their time and other resources. Most cyber defenses have been designed to be optimally strong and effective and prohibit or eliminate attackers as quickly as possible. Our complementary approach is to deploy defenses that seek to maximize the consumption of attackers’ time and other resources while causing as little damage as possible to the victim. This approach is consistent with zero trust and similar mindsets which assume breach. The Sludge Strategy introduces cost-imposing cyber defense by strategically deploying friction for attackers before, during, and after an attack using deception and authentic design features. We present the characteristics of effective sludge and show a continuum from light to heavy sludge. We describe the quantitative and qualitative costs to attackers and offer practical considerations for deploying sludge in practice. Finally, we examine real-world examples of U.S. government operations to frustrate and impose costs on cyber adversaries. We encourage research and further exploration of how sludge can slow attackers.

Keywords: sludge, nudge, cybersecurity, choice architecture, deception, deterrence

1. Introduction

In their 2009 book, Richard Thaler and Cass Sunstein introduced the concept of “nudge,” which they defined as an intervention that “alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives” (Thaler & Sunstein, 2009). Nudges differ from mandates or “shoves” by providing opportunities for those who are nudged to opt out of the nudge with minimal effort or cost. Nudges can be effective if those being nudged have no strong preferences about the focus of the nudge, assume that the nudged behavior is a social norm, or are unaware of the nudge.

Thaler advocates for policymakers to “nudge for good.” This concept has been remarkably popular and has been applied in such diverse areas as registration for organ donation, contributions to retirement accounts, and consumption of healthy food. Recently, proposals have been made to apply nudges to cybersecurity. For instance, password meters encourage users to select strong passwords.

In a later edition of their book (and a separate book by Sunstein (2021)), Thaler and Sunstein (2021) introduced a contrary concept: sludge, defined as “any aspect of choice architecture consisting of friction that makes it harder for people to obtain an outcome that will make them better off.” Sunstein notes that sludge takes many forms, including excessive wait times, unjustified overhead, and “dreary” requirements. Sludges differ from nudges by making it difficult, if not impossible, to opt out of the sludge. Sludges raise transaction costs and make it difficult for people to reach their legitimate goals. Classic examples of sludge are: streaming services that offer free or low-cost subscriptions that automatically convert to much higher fees at the end of the trial period and which require subscribers to

go through laborious steps to cancel; and mail-in vouchers that require customers to provide extensive documentation (much of which the customers rarely retain). Thaler also cites phishing warnings as an example of sludge for *legitimate* users. When a user mostly receives emails from external senders, a warning that the email is from an external sender becomes nothing more than visual clutter that wastes time and attention—unnecessary friction that distracts from the goal of reading and responding to email, without much benefit to the purported benefit of greater security.

The kinds of sludge Thaler and Sunstein consider have largely been “bad” sludge from the perspective of a decision maker. In this paper, we propose sludge for *good* in cybersecurity. That is, we discuss sludges that are good for defenders and designed to slow down cyber attackers. In this sense, sludges for good are a subset of deterrence, which is typically defined as preventing an adversary from taking unwanted actions. Sludges differ from penalties, coercion, compulsion, and other forms of deterrence, but may be more effective in repelling those who wish to harm others. Sludges provide another tool for decision-makers to use in repelling adversaries. By influencing the choices of adversaries, or slowing them down, sludges can deter, prevent, or raise the cost to adversaries in achieving their goals.

Put another way, slowing attackers raises the opportunity cost for attackers. Conducting a successful attack is not free for attackers. At the very least they require time to choose targets, conduct the attack, and determine when to abandon the attack. From the attacker’s perspective, dealing with sludge costs them money, time, and effort. Attackers also give up any gains they might have acquired if they had chosen other attacks or other victims.

Deliberately influencing attackers’ decisions and behavior is an emerging idea. U.S. Cyber Command talks about “imposing cost” on adversaries (Nakasone, 2020). This view implies offensive consequences and retaliation for bad behavior but could be expanded to include defensive tactics that make it more difficult and expensive for adversaries to conduct cyber attacks. Attackers expect their campaigns to generate a positive return on investment and will consider what costs are appropriate relative to the benefits of achieving a given goal. Budgets and resources are finite so this constrains the spectrum of actions an attacker can execute in a given campaign. Certain kinds of attacks, such as supply chain backdoors, generally require the skills and resources of a nation state.

To date, most cyber defenses have been designed to be optimally strong and effective and prohibit or eliminate attackers as quickly as possible. We propose

a complementary approach, which is to also deploy defenses that seek to maximize the consumption of attackers’ time and other resources while causing as little damage as possible to the victim. This approach is consistent with the “assume compromise” mindset whereby defenders treat all systems as not secure and already compromised. To understand and effectively impose cost by slowing attackers, new theory and doctrine are needed to create a common frame of reference. We present the characteristics of good sludge and provide a continuum from light to heavy options. We conclude with recommendations for practitioners.

2. Nudges and Sludges in Cybersecurity

At a high level, Thaler and Sunstein use the phrase “choice architecture” to describe the design in which choices are presented to people. These choices are made easier by nudges and more difficult by sludge. Nudges gently steer people in a direction that increases welfare, including cybersecurity, and are commonly intended to make good outcomes easy. Traditionally, nudges have been used to encourage well-intentioned users to behave in a way that they are better off for doing so. These choices are not guaranteed, but research shows that they are selected more often.

In cybersecurity, both defenders and attackers make decisions. Research has focused primarily on the impact of nudges on defenders and legitimate users. Researchers in one study, for instance, conducted a literature review of 71 papers on technology-mediated nudges (Caraban et al., 2019). They identified 23 distinct mechanisms of nudging which they grouped into six categories: facilitate, confront, deceive, social influence, fear, and reinforce. Only 9% of the papers reviewed were related to security and privacy, and none used deception to promote a particular outcome for either legitimate users or attackers.

Zimmermann and Renaud established that nudges in cybersecurity have four attributes: “predictability of the influence and outcome, involvement of automatic cognitive processes, equality of choice costs, and retention of all pre-nudge choices” (Zimmermann & Renaud, 2021). They applied these principles to decisions such as choosing more secure public WiFi networks and encrypting a smartphone. Peer et al. found that nudges aimed at average users were less effective than nudges that were personalized by decision-making style (Peer et al., 2020). Based on their calculations, the median time to crack passwords was 4.2 times longer for personalized password nudges than for random assignments or no nudge.

Nudges have also been studied to help software

developers make secure choices. One recent study attempted to nudge developers into using safe code snippets on the web rather than unsafe code (Fischer & Grossklags, 2022). Among 218 participants, those receiving nudges produced more functional and secure code in comparison to the control group.

Because they are optional, nudges are not always successful in achieving a specific outcome. X, for example, encourages two-factor authentication (2FA) but reports that only 2.5% of users have it enabled (Twitter, 2022). In a large-scale, in-the-wild study of 2FA adoption messages on Facebook, researchers found that messaging and design strategies can increase adoption, especially when incorporating personalized prompts (Golla et al., 2021). Nevertheless, 2FA remains an opt-in choice on most platforms. When prompts are too aggressive and persistent, nudges produce negative by-products. Users of Apple products are continually reminded to enable 2FA even though it is optional, resulting in user irritation (Apple Inc, 2017).

Cybersecurity research and practice that is focused on influencing attacker behavior is best characterized as sludge for attackers. Some researchers have considered a defensive goal of delaying attackers, rather than denying them, using decoy systems with packet delays (Landsborough et al., 2021). In a laboratory test, the delay tactic increased the average run time from 243.0 seconds (control) to 687.62 seconds and decreased attacker success, while a denial tactic occupied less time and produced lower attacker success. “This technique,” they write, “provides the appearance of poor network performance which can nudge an attacker into moving to a new target—useful if the attacker’s original target was a valuable or vulnerable system.” Others have examined oppositional human factors as a domain of deception intended to nudge attackers into negative affective states (K. J. Ferguson-Walter, Gutzwiller, et al., 2021). In particular, they found in one penetration experiment with professional red teamers where deception was used that participants exhibited confusion, self-doubt, surprise, and frustration. Other work in deception has continued to examine not only discovering adversary activity or deflecting them, but also in depleting their resources. For example, another study explored deception for malware using honey files and honey credentials (Sajid et al., 2021). The researchers measured 13-15% overhead to the system owner of orchestrating two depletion deception techniques but did not attempt to measure the cost borne by the attacker.

In 2017, Shortridge (2017) proposed a new defensive framework that leverages “learning exploitation,” raising the cost of attack by destabilizing attackers’

ability to learn. Within learning exploitation, this approach included sludge-like interventions that make it harder for attackers to learn information about target systems and that introduce unreliability into attack operations. Bogus credentials, for instance, waste attacker time and attention, slowing down their operations. This work also proposed introducing strategic non-determinism into systems to raise the cost of attack during the reconnaissance phase. For example, a defender could make normal endpoints appear like a different malware analysis sandbox upon each startup by adding “hollow but sketchy-looking artifacts” such as debuggers and virtualization libraries.

Two practitioners have reinvigorated the design of deception environments to add “anticipatory mechanisms that impede the success of [attackers’] operations” (Shortridge & Petrich, 2021). “Understanding how attackers make decisions allows software engineers to exploit the attackers’ brains for improved resilience,” they write. The goal of their proposed deception environments is to “disrupt attackers’ abilities to learn and make decisions,” introducing friction into attack operations that slows down attackers while also allowing defenders to collect information about attacker actions.

3. Cyber Sludge in Operations

Three events over the past four years have illustrated actions consistent with slowing cyber attackers using sludge: defense of the 2020 U.S. elections, responses to Russia’s invasion of Ukraine, and counter-ransomware efforts. Although they have not previously been characterized as sludge, we describe how these examples demonstrate and achieve sludge-like impacts.

Sludge was not inevitable for any of these events. The cybersecurity community in the public and private sectors could have exclusively pursued zero tolerance, complete elimination of the problems using technical and non-technical solutions. Instead, these examples offer support that *slowing* adversaries was a component of the strategy.

With regards to countering adversarial cyber activities, the United States has released public comments primarily around sanctions and outing attacker behavior rather than about specific technical details of online operations. This approach is consistent with the precedent of protecting sources and methods. Revealing operational details such as the use of network throttling, for example, may give attackers knowledge to potentially detect and avoid the sludge. One byproduct is that it is easier for non-government observers to estimate qualitative costs than precise quantitative costs.

The U.S. has spoken generally about its desire to slow and disrupt attackers. In an interview, the National Security Agency's (NSA) Chief of Adversary Defeat said that "What we really want to try to do is aggravate, disrupt the adversary so they can't do the things they want to do—doesn't mean we're going to stop them, right? These are persistent adversaries, but to make it harder, to make them alter their schedule, their approach—make them second guess what they're doing and make sure they know that it's not going to be without some kind of cost" (National Security Agency, 2022). These statements reflect a sludge-like strategy.

Nation-state cyber threats remain a focus for United States national security. "We've got to put sand and friction in [adversary] operations so they don't just get free shots on goal," said Rob Joyce, former Director of the NSA's Cybersecurity Directorate, in 2021 (The Aspen Institute, 2021). Persistent engagement, he explained further, is about more than offensive cyber; releasing information about tools and infrastructure is also successful in slowing adversaries.

3.1. Election Security

The protection of electoral systems against interference and influence is integral to democracy. In the United States this outcome involves the combined efforts of federal, state, local, and private sector partners. Election-related cyber threats have been observed in various forms from misinformation to denial of service attacks.

General Paul Nakasone, former Commander of U.S. Cyber Command, testified before Congress ahead of the 2020 elections that "USCYBERCOM is working with the combatant commands, DHS, FBI, across the Intelligence Community, and in conjunction with private sector and foreign partners to improve understanding and act to contest and frustrate adversary cyber activities" (Nakasone, 2019). Frustration is one byproduct of sludge when friction impedes an adversary's goal.

Election security is not limited to the United States. Before the 2017 French presidential election, Emmanuel Macron's campaign team deliberately created false email accounts and fake documents (Nossiter et al., 2017). The stated goal of this deception was to slow down Russian attackers. When gigabytes of stolen data from Macron's campaign were released online, it included real and forged emails. Despite speculation in the press about the effectiveness of the deception on the Russians, no official analysis was released.

3.2. Russia-Ukraine Crisis

The United States, together with dozens of other countries, imposed numerous sanctions against Russia in response to their military attack against Ukraine in early 2022. U.S. officials have reported that the financial sanctions successfully imposed cost, both monetary and psychological, on slowing some Russian cyber attacks. "We've definitively seen the criminal actors in Russia complain that the functions of sanctions and the distance of their ability to use credit cards and other payment methods to get Western infrastructure to run these [ransomware] attacks have become much more difficult," Rob Joyce told The Cipher Brief (2022).

USCYBERCOM, in partnership with the Security Service of Ukraine, also revealed malware used against Ukraine in order to disrupt cyber attacks (U.S. Cyber Command, 2022). This form of outing allows for increased detection of malicious activity and thus imposes friction on the attackers who otherwise benefit from being undetected. While the release did not make a public attribution of the threat actor, it still may have slowed and disrupted the actors' activity.

3.3. Ransomware

Ransomware has become a serious and elusive cyber threat. Ransomware attacks rose during the COVID-19 pandemic when victims included healthcare, financial services, and government systems. Researchers and practitioners have proposed various technical countermeasures, including ransomware honeypots and honeyfiles (Beaman et al., 2021). Nevertheless, attacks remain persistent.

The challenge of ransomware is not simply technical, but also because of the ability for criminals to profit from it. Safe harbor in some nation-states limits the ability for criminal prosecution. However, international financial transactions are essential for ransomware payments and the more friction to financial benefits, the higher the opportunity cost for attackers.

Rob Joyce reported in May 2022 that ransomware activity had declined in the early months of 2022, a trend he attributed, in part, to sanctions against Russia making it more difficult for attackers to buy infrastructure and transfer money (CYBERUK ONLINE, 2022). Thus, sludge had a desirable effect on slowing ransomware.

4. Sludge Strategy for Slowing Attackers

We propose a Sludge Strategy for cyber defense that prioritizes investments into techniques, tools, and technologies that add friction into attacker workflows and raise the cost of conducting operations. Defensive

choice architects can leverage all forms of cost when designing sludge interventions against attackers. To aid in the adoption of weaponized choice architecture, we developed a strategy to help defenders understand the spectrum of potential sludge interventions to deploy against attackers.

Table 1 illustrates how selected defensive techniques offer degrees of sludge for attackers. Each row contains a defensive technique and describes the cost(s) imposed on an attacker. Light sludge describes effects that produce low friction for attackers and heavy sludge produces high friction. The heavier the sludge, the harder it is for attackers to trudge through it. The types of sludge and types of cost in the Table are described in the following sections.

4.1. Quantitative and Qualitative Costs to Attackers

The costs borne by attackers are both quantitative and qualitative. Table 1 shows two types of quantitative cost (monetary and time) and two types of qualitative cost (information and psychological). This section presents relevant aspects of each type and some illustrative examples.

Quantitative Costs. Most attackers must spend money to develop or purchase tools and technologies (including exploits and infrastructure) and pay for talent on their teams, which can be highly specialized (exploit developers, operators, etc.). The ransomware operator Trickbot purportedly invested more than \$20 million into “infrastructure and growth of their organization” in 2021 alone, including investment in technology, human capital, communications, software development, and extortion activities (Burgess, 2022a). The LAPSUS\$ group, an amateur cybercriminal organization, attempted to recruit insiders by offering \$20,000 per week to employees willing to hand over their remote access credentials. Nation states may apply even more investment than cybercriminal organizations, especially in specialized skill sets. By one estimate, the Stuxnet attack costs the offense \$300 million (Slayton, 2016). Attack operations require careful budgeting across a variety of activities.

Attackers must consume time to conduct their operations, which is in limited supply. Each unit of time they expend on one activity cannot be spent on another. Cybercriminals may have target revenue goals within a given quarter or year, and nation-state attackers may have mission goals within a specific time frame, too. If enough time passes without successful compromise, the business or mission will suffer—or be abandoned in favor of a new endeavor. Time costs can be qualitative

as well. Even cybercriminals are conscious of burnout if they continuously work long hours (Burgess, 2022b).

Qualitative Costs. The qualitative costs of sludge should not be underestimated and can include information, psychological, and reputational costs (Sunstein, 2020).

Attackers require information about targets and their systems to be successful; collecting information can be a form of sludge if it takes effort to acquire. In order to send a spearphishing email an attacker must know the victim’s address. The more difficult it is to find, process, or evaluate this information, the more friction is imposed. Conversely, attackers are also susceptible to information overload. Excess information impedes the decision-making process, resulting in a poor decision or decision paralysis.

Attackers, being human, experience negative psychological impacts such as feeling frustrated, dismayed, ashamed, inferior, confused, helpless, stressed, worried, or discouraged (K. J. Ferguson-Walter, Gutzwiller, et al., 2021). Not only can sludge induce these effects, but the effects are sludge to attackers because they impose undesirable friction to achieving their objectives. Some validated scales do exist to measure psychological impact, but they require interaction with the subject which is seldom available from cyber attackers. The Cyber Operations Stress Survey, for instance, uses self-reported measures of fatigue and cognitive workload (Dykstra & Paul, 2018). However, the developers of this instrument found that cyber operations longer than five hours had significant effects on fatigue and frustration and nearly all cognitive workload factors. New approaches will be required to passively infer the effectiveness of imposing psychological cost on attackers.

There is insufficient insight and scientific study about the possible reputational costs to attackers. In theory, successful attacks could raise threat actors’ reputation and unsuccessful or leaked activity could lower reputation. In public relations, metrics include sentiment analysis, stakeholder surveys, and opinion polls; these do not appear to be used to measure threat actor reputation today.

4.2. Types of Sludge

Sludge can be implemented by defenders using both authentic design features and deception, as shown in Table 1. From the perspective of system owners, authentic features are those that are native to products and utilized by end users and administrators and used to protect their own use of a system. Legitimate system

| | Type ¹ | Light Sludge ² | Medium Sludge ² | Heavy Sludge ² |
|--|-------------------|---------------------------|----------------------------|---------------------------|
| Login Banners | A | M | | |
| Authentication | A | I, T | | |
| Decompression Bombs | D | | M, T | |
| Network Throttling | D | | P, T | |
| Perception of Deception | D | | P, T | |
| Immutable and ephemeral infrastructure | D | | I, M, P, T | |
| Deception Environment | D | | | I, T, P |
| Outing Tools / Infrastructure | A | | | M, P |
| Public Attribution/ Outing Attackers | A | | | P |
| Sanctions | A | | | M, P |

¹ A: Authentic Design Feature, D: Deception

² I: Information Cost, M: Monetary Cost, P: Psychological Cost, T: Time Cost

Table 1. Examples of light, medium, and heavy sludge that impose friction on cyber attackers, including the type of sludge and four types of cost to attackers.

users, for instance, have passwords to authenticate themselves that can also impose friction on an attacker who wishes to access a target. Deception—by creating fictitious data including fake accounts, database records, files, and systems—can also impose friction for attackers, since accessing that data can alert an administrator or invoke a defensive response. The names of these fictitious items traditionally carry the prefix “honey-,” as in honeypot or honeytokens, named for their ability to attract and trap attackers.

Authentic Design Features. System owners commonly employ warnings, notices, and other banners before users log in to notify users of acceptable use. The U.S. Department of Defense, for instance, mandates a standard notice and consent on all systems (STIG Viewer, 2015). Users must acknowledge such messages before gaining access. These messages may also seek to deter unauthorized users as a “no trespassing” sign for fear of monitoring and prosecution. It is unlikely that banners impose friction on attackers, but their existence does establish terms of unauthorized access.

Authentication can be an example of sludge for attackers. Passwords, pins, biometrics, and other authentication mechanisms are an intentional barrier to keeping unauthorized users out of accounts and services. In the best case, authentication is a minor inconvenience to legitimate users and a high cost to an adversary. Other examples of authentication sludge for attackers include login push notifications, which validate login requests by notifying an associated mobile device, and periodic key rotation, a recommended practice to limit the number of messages encrypted with the same key which helps prevent cryptanalysis attacks.

Defenders have a distinct advantage over attackers

in the ability to control system accessibility, speed, and responsiveness. For example, the owner of a system can limit the number of login attempts before forcing a discretionary account lockout period. Network throttling allows a network owner to slow down a suspected attacker or aggressive user by limiting the communication speed of data flowing in or out. Access sludge can also be imposed if systems only accept connections for predefined, trusted IP addresses.

On a software level, an authentic feature that produces sludge for attackers is code obfuscation. This approach increases the time, psychological, and information costs necessary for attackers to discover vulnerabilities. Software developers can take steps to make reverse engineering and vulnerability discovery more difficult and time consuming, such as symbol stripping and anti-debugging techniques (Votipka et al., 2020). To our knowledge, the costs imposed by these techniques have not been measured. In recent research, a new mitigation showed promise in injecting delays into the execution of illicit cryptomining on continuous integration (CI) platforms (Li et al., 2022). Their evaluation also showed that this rendered the attack unprofitable and with only small impacts on legitimate CI jobs.

Defenders can design their infrastructure to be immutable and ephemeral, as is becoming an emerging trend in private sector defense through the practice of security chaos engineering. Immutability means that once infrastructure is deployed, it cannot be changed. Attackers frequently take advantage of secure shell (SSH) access in servers; but an immutable server can have SSH access disabled by default, given no changes are allowed, cutting off that attack path. Ephemerality

means that infrastructure lives only for a short period of time, usually the duration of executing a task, before terminating. There may be some impact to end users, specifically software engineers who interact with this infrastructure, in that they may need to alter workflows and update design accordingly (such as finding other ways to troubleshoot problems in production if debugging is not allowed). However, this friction is asymmetric in its impact because it much more drastically changes how attackers engage with the target system than it does for software engineers.

In addition to technical capabilities, an authentic feature of cybersecurity is attribution. This response involves a government or private entity publicly naming an actor or nation-state responsible for particular cyber activity. Attribution is friction for adversaries for several reasons. Attribution draws attention to the activity that prepares and informs a broader community to prepare, detect, and evict similar activity in their environments. Few comprehensive analyses exist which measure the costs borne by attackers because of attribution. However, attribution can also lead to tangible economic sanctions and criminal indictments.

Among the heaviest sludge and most friction for attackers are political, economic, and criminal responses, such as sanctions and indictments. While the costs are more transparent, the impact in cybersecurity may be limited (Romanosky & Boudreaux, 2021). These options are available only to nation-states as shown in the examples in Section 3.

Deception. System owners have the upper hand in ground truth. They possess information about their systems that attackers must endeavor to acquire, reflecting an information asymmetry. Attackers must expend effort both in acquiring information and determining its relevance to their operations. System owners can leverage deception to lead attackers to operate based on false assumptions—resulting in wasted financial, time, and cognitive resources—or to foment fear, uncertainty, and doubt in attackers, who must then expend more resources attempting to differentiate the real from the mirage.

An adversary's perception of deception is an effective form of sludge. In a study of 130 professional red teamers, psychological deception appeared to be effective even if the attacker merely believed it may be in use (K. J. Ferguson-Walter, Major, et al., 2021). This approach provides an unusually high return on investment for defenders.

Decompression bombs are a form of deception sludge that impose friction on attackers after they have stolen data. This technique works by enticing an attacker to steal a seemingly valuable file with

a specially-created decoy that requires an excessive amount of time, disk space, or memory for the attacker to decompress. The approach increases the cost imposed compared with traditional non-compressed decoy files.

Honeypots to deceive and distract adversaries have been applied since the 1980s. Many honeypots are implemented as standalone systems to distract attackers from production systems. Few studies have examined the prevalence and success of honeypots on a large scale. In one study, researchers discovered over 19,000 Internet-facing honeypots in 637 autonomous systems (Morishita et al., 2019). This could be an underestimate but given that there are 100,000 autonomous systems on the Internet it likely shows that honeypots are rare. While a honeypot is a lure for attackers to a decoy system, tarpits are security mechanisms that explicitly aim to slow an attacker's progress. One Internet-wide scan found 215,000 IP addresses in 107 networks among 77 autonomous systems exhibiting tarpit-like behavior (Alt et al., 2014).

Related to honeypots, honey-patches are a technique where a defender patches a known vulnerability but adds functionality to mislead attackers into believing that a failed exploitation attempt was successful (Araujo et al., 2014). The attacker interacts with a decoy environment that consumes time.

4.3. Practical Considerations

Creating and deploying sludge will depend on business decisions including financial implementation costs, impact to legitimate users, and degree of transparency of the sludge to attackers. As Table 1 showed, practitioners have a range of choices depending upon their goals and resources. Sludge can even be a by-product of some authentic features such as authentication. However, there are financial and time overhead costs to defenders in the maintenance and sustainability of deception sludge. Minimizing the friction imposed on defense remains an area of active research (Sajid et al., 2021).

Impact to Legitimate Users. System administrators must carefully balance the potential impact on legitimate users when introducing sludge into attacker workflows. Logon banners, for example, are shown as light sludge in Table 1 and while they slow legitimate users to make a thoughtful decision (Fassl et al., 2021), they produce little friction for attackers who bear few consequences of ignoring them. In contrast, a company could allow users to log in only after they physically badge into the building is light sludge for users but heavy friction for attackers. Similarly, imposing network traffic delay costs for remote network connections might slow

attackers without effecting on-site employees.

Relationship to Deterrence. Sludge, and other approaches to impose cost by influencing attackers' decisions, has an effect on nascent efforts to apply deterrence theory into effective cyber applications (Jasper, 2017). However, sludge and deterrence are not synonymous. Sludge creates friction that hinders attackers in achieving their goals or making beneficial decisions, often resulting in inefficiency or frustration. Cyber deterrence, on the other hand, is a broad, strategic approach aimed at discouraging attackers from launching attacks by threatening retaliation or increasing the perceived cost and difficulty of an attack (Morgan, 2010). While sludge can be considered a *form* of deterrence, it specifically targets the decision-making process of attackers by imposing psychological and operational costs that disrupt their ability to conduct effective attacks, distinguishing it from traditional deterrence methods that focus primarily on retaliation or defense. We suggest that retaliation is consistent with deterrence but not with sludge.

A significant change in cyber deterrence is the need to project security and impose a cost should the attacker attempt to compromise a target. While no single approach which affects the attacker's decision calculus may alone be an effective deterrent, the combined effects of the choice architecture determine defenders' overall deterrence posture. Sludge also serves as a distinct deterrence toolkit by encouraging attackers who are frustrated to look elsewhere, sowing doubt on the validity of their access, and providing false appearances of successful compromise.

Finally, research in the quantitative and qualitative assessments of cyber deterrence remains a challenge and active area of investigation for practitioners and researchers (Llopis Sanchez & Lopes Antunes, 2024).

Measuring Success. Measures of effectiveness are an important aspect for evaluation. Some types of sludge produce data that are visible and measurable. For instance, a defender-controlled honeypot can be monitored by its owner. System owners also have insight into who triggers network throttling, when it occurs, and for how long it took effect. On the other hand, other measures of cost—such as attacker frustration—are more difficult to passively observe and measure (K. Ferguson-Walter et al., 2018). One way to measure success is by comparing the impacts on two networks attacked by the same threat actor where one applies sludge and one does not.

Traditional measures in cybersecurity incident response such as mean-time-to-discovery or mean-time-to-remediation are imprecise for sludge. Instead, new measurements will be necessary. These

must be integrated and considered in consort with other defenses to form a composite picture of overall organizational cybersecurity.

5. Limitations and Future Work

First, we acknowledge that, despite point solutions such as honeypots and emerging government examples, there are nascent implementations and evaluations of sludge as a strategy to impose cost on attackers. We hope that our work encourages further exploration and evaluation. Second, the effectiveness of the strategy may benefit from knowledge of individual human and organizational factors of the attackers. Sludge is, after all, a challenge to the human nature of the attacker. If cyber adversaries embody common personalities and traits, then these would enable many kinds of sludge to be effective regardless of individual differences. Still, future research should examine the traits and characteristics of people for whom various sludge is effective. Third, sludge will not stop a dedicated attacker who has enough resources to persist and adapt despite sludge. Sludge implementations will have to respond and evolve to this type of adversary and also as general cybersecurity evolves. Nevertheless, persistent threat actors are still human and vulnerable to human weaknesses and psychological manipulation.

Cybersecurity professionals often seek to minimize their recovery time, failure rates, and lead times. If adversaries behave likewise, sludge may be used to strategically maximize negative results. That is, it may be possible to slow an adversary's recovery time or increase failure rates in capability development or operations. For example, according to a report, "[Stuxnet] generated malfunctions in the centrifuges of the Natanz enrichment plant at random intervals over months, using different errors each time, and rendering them undetectable to the diagnostic systems in the control room" (Greenberg, 2012).

Future research should explore new types of sludge. One idea is the potential value of fracturing adversary teams from within by employing nudges and sludges. Organized attackers commonly rely on teams with specialized roles. Each human element in attack operations is subject to cognitive bias and the natural reticence to admit fault, leading to pointing fingers at each other when things go wrong. This internal division makes it harder to carry out successful operations. Sludge could enable defenders to slow down individual parts of this process.

The Sludge Strategy introduces new cost-imposing cyber defense by strategically deploying friction for attackers. The strategy broadens the options

beyond complete denial and is consistent with modern information defense which assumes system and network compromise. We encourage cyber defenders, military planners, and system designers to consider how sludge can improve cybersecurity in their environments. New implementations of sludge must be developed and evaluated by interdisciplinary teams to ensure that they produce the desired outcomes.

References

- Alt, L., Beverly, R., & Dainotti, A. (2014). Uncovering network tarpits with degreaser. *Proceedings of the 30th Annual Computer Security Applications Conference*, 156–165.
- Apple Inc. (2017, April). How do I stop the two-factor authenticatio... - Apple Community. <https://discussions.apple.com/thread/7923133>
- Araujo, F., Hamlen, K. W., Biedermann, S., & Katzenbeisser, S. (2014). From patches to honey-patches: Lightweight attacker misdirection, deception, and disinformation. *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 942–953.
- Beaman, C., Barkworth, A., Akande, T. D., Hakak, S., & Khan, M. K. (2021). Ransomware: Recent advances, analysis, challenges and future research directions. *Computers & Security*, 111, 102490.
- Burgess, M. (2022a). Inside Trickbot, Russia's notorious Ransomware Gang. *Wired*. <https://www.wired.com/story/trickbot-malware-group-internal-messages/>
- Burgess, M. (2022b). The workaday life of the world's most dangerous ransomware gang. *Wired*. <https://www.wired.com/story/conti-leaks-ransomware-work-life/>
- Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.
- CYBERUK ONLINE. (2022, May). *Plenary 1 : Global response, Global impact: Strategic alignment and collaboration*. <https://www.youtube.com/watch?v=7Ywcj8Jdv7w>
- Dykstra, J., & Paul, C. L. (2018). Cyber Operations Stress Survey (COSS): Studying fatigue, frustration, and cognitive workload in cybersecurity operations. *11th USENIX Workshop on Cyber Security Experimentation and Test*.
- Fassl, M., Gröber, L. T., & Krombholz, K. (2021). Stop the consent theater. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.
- Ferguson-Walter, K., Shade, T., Rogers, A., Trumbo, M. C. S., Nauer, K. S., Divis, K. M., Jones, A., Combs, A., & Abbott, R. G. (2018). *The tularosa study: An experimental design and implementation to quantify the effectiveness of cyber deception*. (tech. rep.). Sandia National Lab, Albuquerque, NM.
- Ferguson-Walter, K. J., Gutzwiller, R. S., Scott, D. D., & Johnson, C. J. (2021). Oppositional human factors in cybersecurity: A preliminary analysis of affective states. *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops*, 153–158.
- Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., & Muhleman, D. H. (2021). Examining the efficacy of decoy-based and psychological cyber deception. *30th USENIX Security Symposium*, 1127–1144.
- Fischer, F., & Grossklags, J. (2022). Nudging software developers toward secure code. *IEEE Security & Privacy*, 20(02), 76–79.
- Golla, M., Ho, G., Lohmus, M., Pulluri, M., & Redmiles, E. M. (2021). Driving 2FA Adoption at Scale: Optimizing Two-FactorAuthentication Notification Design Patterns. *30th USENIX Security Symposium*, 109–126.
- Greenberg, A. (2012, June). What Stuxnet's Exposure As An American Weapon Means For Cyberwar. <https://www.forbes.com/sites/andygreenberg/2012/06/01/what-stuxnets-exposure-as-an-american-weapon-means-for-cyberwar/>
- Jasper, S. (2017). *Strategic cyber deterrence: The active cyber defense option*. Rowman & Littlefield.
- Landsborough, J., Carpenter, L., Coronado, B., Fugate, S., Ferguson-Walter, K., & Van Bruggen, D. (2021). Towards self-adaptive cyber deception for defense. *HICSS*, 1–10.
- Li, Z., Liu, W., Chen, H., Wang, X., Liao, X., Xing, L., Zha, M., Jin, H., & Zou, D. (2022). Robbery on devops: Understanding and mitigating illicit cryptomining on continuous integration service platforms. *2022 IEEE Symposium on Security and Privacy. IEEE Computer Society, Los Alamitos, CA, USA*, 363–378.

- Llopis Sanchez, S., & Lopes Antunes, D. (2024). Operation assessment in cyberspace: Understanding the effects of cyber deception. *Proceedings of the 19th International Conference on Availability, Reliability and Security*, 1–8.
- Morgan, P. M. (2010). Applicability of traditional deterrence concepts and theory to the cyber realm. *Proceedings of a workshop on deterring cyberattacks: Informing strategies and developing options for US policy*, 56.
- Morishita, S., Hoizumi, T., Ueno, W., Tanabe, R., Gañán, C., van Eeten, M. J., Yoshioka, K., & Matsumoto, T. (2019). Detect me if you... oh wait. an internet-wide view of self-revealing honeypots. *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 134–143.
- Nakasone, P. M. (2019, February). Statement of General Paul M. Nakasone Commander United States Cyber Command Before the Senate Committee on Armed Services. https://www.armed-services.senate.gov/imo/media/doc/Nakasone_02-14-19.pdf
- Nakasone, P. M. (2020). Operations in cyberspace and building cyber capabilities across the department of defense. *117th Cong., 2020. (Testimony of General Paul M. Nakasone)*.
- National Security Agency. (2022, September). *Look around: Women in cybersecurity episode 6*. <https://www.youtube.com/watch?v=5LpJIORWaYE>
- Nossiter, A., Sanger, D. E., & Perlroth, N. (2017). Hackers came, but the french were prepared. *New York Times*, 9.
- Peer, E., Egelman, S., Harbach, M., Malkin, N., Mathur, A., & Frik, A. (2020). Nudge me right: Personalizing online security nudges to people's decision-making styles. *Computers in Human Behavior*, 109, 106347.
- Romanosky, S., & Boudreaux, B. (2021). Private-sector attribution of cyber incidents: benefits and risks to the US Government. *International Journal of Intelligence and CounterIntelligence*, 34(3), 463–493.
- Sajid, M. S. I., Wei, J., Abdeen, B., Al-Shaer, E., Islam, M. M., Diong, W., & Khan, L. (2021). SODA: A System for Cyber Deception Orchestration and Automation. *Annual Computer Security Applications Conference*, 675–689.
- Shortridge, K. (2017). Big game theory hunting. *Black Hat Briefings USA*. <https://www.youtube.com/watch?v=CqwzWoJdbTc>
- Shortridge, K., & Petrich, R. (2021). Lamboozling attackers: A new generation of deception: Software engineering teams can exploit attackers' human nature by building deception environments. *Queue*, 19(5), 26–59.
- Slayton, R. (2016). What is the cyber offense-defense balance? conceptions, causes, and assessment. *International Security*, 41(3), 72–109.
- STIG Viewer. (2015, June). The operating system must display the standard mandatory dod notice and consent banner before granting local or remote access to the system. <https://www.stigviewer.com/stig/general-purpose-operating-system-srg/2015-06-26/finding/V-56585>
- Sunstein, C. R. (2020). Sludge audits. *Behavioural Public Policy*, 1–20.
- Sunstein, C. R. (2021). *Sludge: What stops us from getting things done and what to do about it*. MIT Press.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin Books.
- Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition*. Penguin Books.
- The Aspen Institute. (2021, September). *2021 Aspen Cyber Summit: DAY 1*. <https://www.youtube.com/watch?v=ww-CxxEj8Sc&t=12086s>
- The Cipher Brief. (2022, July). *View from the NSA with Rob Joyce, Director of Cybersecurity — Cyber Initiatives Group*. <https://www.youtube.com/watch?v=e-Sko0Kersc>
- Twitter. (2022, January). Account security - twitter transparency center. <https://transparency.twitter.com/en/reports/account-security.html>
- U.S. Cyber Command. (2022, July). *Cyber National Mission Force discloses IOCs from Ukrainian networks*. <https://www.cybercom.mil/Media/News/Article/3098856/cyber-national-mission-force-discloses-iocs-from-ukrainian-networks/>
- Votipka, D., Rabin, S., Micinski, K., Foster, J. S., & Mazurek, M. L. (2020). An observational investigation of reverse engineers' processes. *29th USENIX Security Symposium*, 1875–1892.
- Zimmermann, V., & Renaud, K. (2021). The nudge puzzle: Matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction*, 28(1), 1–45.