Clandestine Distortion and Sanctions

Jelena Vićić[1] Jason Davis[2] Rupal N. Mehta[3]

December 9, 2025

[1] Indiana University Bloomington, Co-First Author
[2] Florida State University, Co-First Author
[3] University of Nebraska-Lincoln

Abstract

Inspired by the Stuxnet virus cyber operation and other developments in cyberwarfare, we present and analyze a formal model of *clandestine distortion*. Specifically, we explore a scenario where a sending state can generate uncertainty in an adversary's understanding of their own capabilities by launching an unobserved attack that degrades the adversary's signal of their own probability of success in achieving some policy goal that the sending state would rather they abandon - in the Stuxnet case, this was Iran's efforts to develop a successful nuclear program. We show both that (1) clandestine distortionary attacks can successfully lead a state to abandon a policy goal by complementing overt forms of coercion (e.g. sanctions); (2) the effectiveness of this class of clandestine attack quickly degrades as the adversary becomes more confident in the sending state's capability of launching such an attack. In the limit, clandestine distortionary attacks provide no benefit to the sending state, but the sending state needs to continue launching them despite their costs to prevent the adversary from becoming even more emboldened than they would otherwise be in pursuit of their goal. We illustrate the implications of the model through a discussion of the Stuxnet cyber operation and subsequent negotiations over Iran's nuclear program, as well as US-led "left-of-launch" operations against North Korea.

Introduction

Starting sometime around 2008, scientists working at Iran's Natanz nuclear facility were beginning to notice a problem. While some centrifuges used for uranium enrichment would be expected to fail in the course of regular operations - up to 10% by some estimates (Zetter, 2015*a*, p.1) - the failure rate at Natanz was significantly higher, perhaps more than 20% (Zetter, 2015*a*, p.3). Centrifuges regularly spun out of control and exploded, while the instruments used to monitor the centrifuges showed no sign of problems that would explain these failure (Sanger, 2018, p.28).

Was the root cause incompetent engineering? Bad parts? Sabotage? Neither the scientists involved nor Iranian officials had any idea, prompting recriminations, temporary shutdowns of large blocks of centrifuges, and firings of nuclear scientists. (Sanger, 2012) Doubts began to arise about the ability of Iranian scientists and engineers to actually deliver on the project of a functioning nuclear program.

We now know that it *was* sabotage: specifically, these problems were caused by the Stuxnet virus, which was the core part of a covert cyber operation codenamed "Olympic Games" (Lindsay, 2013; Rid and Buchanan, 2015). The self-replicating virus had been *specifically designed* to go undetected (Falliere and Chien, 2011), and make it seem like Iranian scientists and engineers were incompetent; as one of the participants in the program put it, "[t]he thinking was that the Iranians would blame bad parts, or bad engineering, or just incompetence. The intent was that the failures should make them feel they were stupid, which is what happened." (Sanger, 2012)

In this paper, we explore the strategic implications of this kind of attack using a game theoretic model. Specifically, we look to explore the use of a certain class of operation: clandestine attacks that are designed to impact the perceived likelihood of success of some target project while being conducted without attribution *or detection*. This kind of operation has become more salient with the advancing importance of strategic competition in cyberspace

(Fischerkeller, Goldman and Harknett, 2022), and the paper discusses illustrative examples from this domain, but it is not *unique* to cyber conflict. Our model can provide useful insight into any case of sabotage that shares these core features, many of which, by virtue of being clandestine, we may never observe.

The model explores the utility of this kind of operation as a tool of coercion. Can clandestine distortionary attacks contribute to an overall strategy that aims to compel an adversary to abandon pursuit of a project? The results of the model have important implications for statecraft: while clandestine distortions *can* meaningfully advance a strategy of coercion, this ability is attenuated significantly by an adversary's *expectation* of such attacks. As an adversary becomes more confident in the attacker's ability and willingness to use such capabilities, they adjust their estimates of success in a project accordingly, taking into account the possibility that their perceived failures are the result of foreign interference. In the limit, this can produce equilibria in which this clandestine activity produces *no benefit* relative to equilibria where the tool did not exist, but where attacking states need to continue to use it anyways lest the adversary become emboldened even further in the pursuit of their project.

The results suggest strong strategic reasons for keeping capabilities secret and have important implications for evaluating the use of tools when their use might reveal such capabilities. Moreover, the model outlines a mechanism by which clandestine action can contribute to coercion, contrasting with theoretical arguments outlining the limitations facing coercion in this space (Schelling, 1960; Borghard and Lonergan, 2017). The paper thus contributes to our understanding of cyberconflict, and of coercion via covert action more broadly.

Scope Conditions

What kinds of situations is this paper's model useful for understanding? There is no question that the scope for the kind of attack described in this paper has advanced exponentially as cyber-technology has progressed. The key features of an attack that we are looking for is that it be *clandestine* (i.e. covert *and* undetected), and that it *distort* (i.e. that it degrade the signal received of likely success in some project) - these features are much easier to achieve in cyberspace than they are in traditional, physical intelligence operations, and this is likely to become increasingly true as rapid developments in generative artificial intelligence (AI) continue, in ways we are only beginning to grapple with.

Such attacks may *also* produce physical damage or degradation of an adversary's capacities - indeed, this was the case with Stuxnet, though its overall physical impact on progress was arguably fairly limited (Lindsay, 2013) - but the key mechanism we are interested in exploring in this paper is their impact through informational channels.

Stuxnet is an especially useful example to explore through the lens of this model because it was, *largely by accident*, ultimately revealed for our observation. The United States and Israel had spent billions of dollars trying to prevent this occurrence - even including code in the virus that would result in it stopping and deleting itself in 2012 to try to prevent its discovery - but the virus accidentally "escaped containment" to computers outside Iran because of a coding error, and then was inadvertently discovered and tracked to Iran by computer security experts in Belarus (Zetter, 2015*a*, p.5). We might expect that *most* operations that this model would be usefully applied to are not observable (Oppenheimer, 2024).

However, it remains important to emphasize that attacks with similar features can be seen going back decades: Soviet "peacetime" sabotage, for instance, often targeted bridges and infrastructure in ways designed to make it look as though problems occurred by accident, attempting to degrade US allies' understanding of their capabilities of resisting Soviet threats and incursions (Kelton, 2025). Similarly, Soviet "illegals" operations entailed embedding deep-cover agents in order to, in many cases, feed misinformation that would not be detected but which could alter perceptions of the expected value of pursuing certain policy choices over others (Riehle, 2020).

Stretching the conceptual apparatus of the model only slightly, we can also observe similar features in United States subversion efforts in Latin America.[1] The US has often been accused of attempting to clandestinely distort adversarial governments' signals of the stability of their governments in the face of revolutionary threats by, for instance, fomenting or staging protests; the US might do this in order to persuade a government to abandon its goal of retaining power in favor of fleeing for personal safety, or making other concessions to the US in exchange for actions that might promote greater stability (Nutt, 2025; Schrader, 2018) These features align well with the mechanisms our paper's model describes.

We are thus confident that while this paper's model will have increasing purchase as strategic cyber competition becomes an even more prominent feature of the global threat environment, the mechanisms outlined have even broader implications for understanding both historical events and operations, as well as the likely trajectory of the future.

Overt and Covert Coercion

Political scientists have long sought to characterize the conditions under which states can coerce adversaries to adopt their preferred polices. Scholars of economic coercion (e.g. most prominently, but not exclusively, sanctions) have explored a wide variety of characteristics that might condition when coercion is effective: these include regime type (Peksen, 2019), expectations of future conflict (Drezner, 1999), the number of parties involved in sanctions (Bapat and Morgan, 2009), the support of the target's major trading partners (McLean and Whang, 2014), the export profiles of the target (Kavakli, Chatagnier and Hatipoğlu, 2020), the incentives of a state to enforce sanctions given the cost inflicted on domestic firms (Ba-

[1] We thank Kolby Hanson for this observation.

pat et al., 2024), domestic political structures (Allen, 2005, 2008), or the degree to which sanctions are targeted at elites (Morgan and Schwebach, 1995; Brooks, 2002; Lektzian and Souva, 2007; Davis, 2025).

Other work on sanctions has explored potential uses of sanctions *beyond* coercion: for instance, sanctions may be imposed in an effort to spark regime change (Marinov, 2005), to hamper a target's future capabilities (McCormack and Pascoe, 2017; Kustra, 2023; Joseph, 2023), or to credibly signal resolve (Lektzian and Sprecher, 2007). However, recent work has also demonstrated how these different mechanisms, while conceptually distinct, can have important complementarities with each other: for instance, increasing the probability of regime change can make achieving coercion more likely (Davis, 2025), and constraining an adversary may in some cases enhance the efficacy of coercion (Di Lonardo and Tyson, 2024). Similarly, in this paper, we explore how a particular, novel mechanism of action - *clandestine distortion* - can complement better understood, overt mechanisms of coercion.

The majority of work on coercion has focused on *overt* actions, in large part because, going back to Schelling (1960), the conditions under which coercion has been expected to be effective are almost definitionally absent from cases of covert action (Poznansky and Perkoski, 2018): namely, that a target must be aware that a cost is being imposed on them because of a particular behavior, and that this cost will be removed if they change this behavior. Borghard and Lonergan (2017) explore the degree to which this classic coercive logic applies in cyberspace, and largely conclude that coercion in cyberspace is less likely to be effective because of how attacks in this domain diverge from these characteristics. Baliga, Bueno de Mesquita and Wolitzky (2020) explore deterrence - a form of coercion - when attribution of attacks is *imperfect*, but in their model the nature of the coercive threat is actually quite overt and direct: the defender openly attempts to coerce imperfectly attributed attackers.

When then - if ever - is covert coercion possible? Jun (2021) explores the particular case of ransomware, where the threat and conditional punishment are clear by the very nature of the attack. Joseph (2025) explores a model in which attacks are unattributed, but where there are several potential attackers, resulting essentially in distributed attribution for attacks, and effective coercion when the defending state begins to recognize patterns where particular behaviors are consistently followed with attacks, in almost a "Pavlov's dog" approach to coercion.

The case we explore in our paper is an even harder case, however, since the attacks are not only covert (unattributable) but clandestine (unobserved). This would seem to violate the basic tenets of coercive logic: if a state does not observe that they are being attacked, how can they know what actions they need to take in order to prevent such attacks? We demonstrate that clandestine distortionary attacks can *still* result in coercion by degrading the expected benefit side of a cost-benefit calculation posed by overt attempts to coerce like sanctions, and indeed our mechanism *relies* on the attack remaining hidden in order to generate the uncertainty that impacts this calculation. The *clandestine distortion* mechanism we propose can successfully achieve coercion so long as the recipient pays *some cost* to pursuing the project that can be recouped by abandoning that project; sanctions are a natural fit as the most likely source of this cost, but the mechanism can apply even more broadly.

Considering Cyber Operations

Similar to sanctions, covert operations may seek to induce a change in a foreign state. However, covert operations aim to do so in a plausibly deniable manner (Cormac, Walton and Van Puyvelde, 2022; Scott, 2004). Covert operations can take various forms, including propaganda, political and economic action, paramilitary operations, lethal action (*1947 National Security Act*, 1947) and others. While often seen as a controversial tool, covert action carries an attractive benefit of plausible deniability (Joseph and Poznansky, 2018; Cormac and Aldrich, 2018) and can be driven by desire to limit escalation (Carson, 2018). It is worth making a distinction between covert and clandestine action. According to the U.S. Congress definitions, the former includes activity to "influence political, economic, or military conditions," where the role of the sponsor "will not be apparent or acknowledged publicly," DeVine (2022). Clandestine operations, on the other hand, are defined by the U.S. Department of Defense as "sponsored or conducted by governmental departments or agencies in such a way as to assure secrecy or concealment," (of Staff, 2021). In other words, in the case of clandestine action, both the initiator and the action itself are secret. With the introduction of cyberspace as an environment of exploitation and competition among states (Fischerkeller, Goldman and Harknett, 2022), activities that constitute "secret statecraft" (Lindsay, 2020) have expanded into the new environment.

Cyberspace, famously described as "consensual hallucination" by William Gibson who first used the term (Gibson, 1984), can be defined as an "interdependent network of information technology infrastructures that includes the Internet, telecommunications networks, computer systems, and embedded processors and controllers in critical industries"(Ross and McQuaid, 2021). Having over the years gained the status of a military domain alongside air, land, sea, and space, cyberspace has become an environment where states secretly compete for influence.

Cyberspace is particularly well-suited for covert operations, given that it easily facilitates clandestine action through attribution challenges (Rid and Buchanan, 2015; Egloff and Smeets, 2021; Gartzke and Lindsay, 2015). In other words, "[o]n the Internet, nobody knows you're a dog," (Steiner, 1993). Over the last two decades, we have seen a significant increase in the use of cyberspace to pursue a variety of goals, from stealing secrets to sabotage, with varying levels of success (Valeriano, Jensen and Maness, 2018). From election interference (Vićić and Gartzke, 2024; Vićić and Harknett, 2024) to air defense radar jamming (CCDCOE, 2007), cyber capabilities represent a diverse toolkit that states have been using profusely. As such, it can be easily integrated and used with other foreign policy tools to induce behavioral changes in target states.

Cyber operations often involve intrusion or unauthorized access, and external control of affected networks. Overall, the range of possible effects spans from website defacement and enabling influence operations, which generally produce least harm, all the way to disabling critical infrastructure and interfering with nuclear command and control, which carries most harm (Maness, 2022). In addition to that, these operations do not necessarily need to include harmful action, and may simply be used as a surveillance tool. The span of operational effects and strategic ends that can be pursued through cyberspace is broad and states have been experimenting.

Sometimes contextualized as hassling (Schram, 2021), the effects of cyber operations have often eluded scholars and policy-makers. Research on the effects of cyber operations reveals a nuanced picture: while many once heralded offensive cyber capabilities as game-changing tools of state competition, empirical work increasingly suggests that their strategic impact is constrained and context-dependent. Although states may have used cyber capabilities to achieve strategic effects (Fischerkeller, Goldman and Harknett, 2022; Harknett and Smeets, 2022), studies of coercion and escalation argue that cyber operations rarely achieve effects on their own (Borghard and Lonergan, 2017) and are considered poor coercive tools (Valeriano, Jensen and Maness, 2018). Empirical analysis of conflicts in Ukraine and Syria finds little evidence that cyber operations shape battlefield outcomes (Kostyuk and Zhukov, 2019; Maschmeyer, 2024), while the analysis of high-profile incidents such as Stuxnet indicates that large-scale and lasting effects are difficult to sustain (Lindsay, 2013). Overall, skepticism abounds in scholarship on the effectiveness of cyber operations, suggesting that such activities may be more suitable for espionage and shaping the information environment (Buchanan, 2020) rather than imposing decisive coercion or physical damage. Against such background, it is puzzling why states persist in their cyber activities.

The burgeoning literature on the psychological and cognitive effects of cyber operations underscores that beyond technological breaches, cyber incidents have meaningful human-level impact on emotions, attitudes, and cognition (Shandler and Canetti, 2024). Experimental work shows that "seemingly inconsequential" cyberattacks may cause psychological distress (Shandler, Gross and Canetti, 2023) and depending on the incident lethality, may influence civilian support for retaliation (Shandler et al., 2022) as well as affect political attitudes on questions of security, civil liberties, and military retaliation (Gross, Canetti and Vashdi, 2017). Other work (Jardine, Porter and Shandler, 2024) examines how uncertainty about cyber operations affects public opinion, decision-making, trust, and perception of threats. Collectively, extant research shows that beyond technical and economic costs, cyber operations also generate psychological and cognitive ripple-effects.

Theorizing Uncertainty: Second-Order Effect of Cyber Operations

Uncertainty in international relations is usually conceptualized as uncertainty about others – their capabilities (Blainey, 1988; Spaniel, 2022), intentions, and resolve (Jervis, 2017). In the context of rationalist explanations for war (Fearon, 1995), uncertainty is theorized as *sine qua non* of conflict (Gartzke, 1999). War in this view is the result of an inherent uncertainty about the world and the adversary. Rarely, if ever, is one uncertain about oneself, as is she about others in the international system.[2] In this paper, we theorize that cyber operations, due to their core features, allow state actors to induce in their competitors the cognitive effect of uncertainty about one's own capabilities.

Cyber operations may unfold in complete secrecy - intrusions, exploitations, and system compromises can remain entirely invisible to the target. More so than many other instruments of statecraft, cyber capabilities are able to achieve a wide range of effects clandestinely; operators often choose them precisely for that secrecy (Lindsay and Gartzke, 2018). In such cases, not only is the identity of the perpetrator concealed, but the action itself may

[2] Davis (2021) is one exception, and explores how in a model of strategic information transfer between the bureaucracy and the leader of the state, uncertainty may be generated about a state's own capabilities.

remain hidden. This form of sabotage generates profound uncertainty - "[e]ven after targets realize that something is wrong, they often remain uncertain about the cause," (Rovner, Cormac and Maschmeyer, 2025). The uncertainty induced by cyber sabotage may create operational friction while compelling the targeted to attribute the problem to internal failings rather than external interference. This represents a second order effect of cyber activity: while first order effects result directly from the technical action, second order effects concern the human response to the perception that the system is not functioning as expected.

Instead of attributing system failures to an external adversary, the target may internalize blame, assuming that the fault lies within its own organization. When the affected system is part of a critical technological enterprise, such as the development of a weapons platform or other strategic capability, this misattribution can have profound second-order consequences. The target's confidence in its technical competence may erode, as leadership begins to doubt the ability of its scientific and engineering community to achieve stated objectives. What begins as a technical disruption thus evolves into a cognitive and organizational one.

From a psychological and political standpoint, this dynamic represents a form of cognitive effect: rather than compelling the adversary through overt force or explicit threat, the operation induces internal doubt, mistrust, and fear. The initiator is not using force overtly to stop the target from completing their intended program. Instead, the initiator is introducing uncertainty into the target's operational calculations and analysis. The expert community may become suspect, their expertise questioned, and the legitimacy of their advice undermined. Alternatively, blame may shift from incompetence to disloyalty, as political leaders come to suspect the presence of internal saboteurs. Such suspicions can fuel paranoia within elite circles, corrode institutional cohesion, and distort decision-making processes.

Conceptually, these are second-order effects of cyber operations - emergent psychological and organizational consequences that arise not from direct physical or digital disruption, but from the way actors perceive and interpret anomalous system behavior. In this sense, clandestine sabotage functions as a subtle form of coercion or strategic manipulation, shaping the adversary's cognition and behavior through uncertainty rather than through overt confrontation.

The kinds of capabilities required to achieve the effects discussed here are often extraordinarily complex and costly to develop. They demand not only advanced technical expertise but also extensive intelligence collection, operational access, and precise timing to ensure that the disruption remains effective.

And yet, the effects that these operations can produce - particularly their psychological and organizational consequences, may exceed their immediate tactical value. A single, well-executed cyber operation that introduces confusion, mistrust, or internal blame within a rival's strategic program can achieve what conventional strikes cannot: the corrosion of institutional confidence and the erosion of decision-making coherence.

Clandestine cyber activity is not, by design, a device for signaling intent or demands. When a cyber operation succeeds in remaining hidden, the target cannot observe the action and therefore cannot reasonably construe it as a communicative act. Indeed, that absence of observability is sometimes precisely why a particular cyber tool is favored.

Signaling may instead occur through other, more overt instruments of statecraft that accompany or follow the clandestine activity. Economic sanctions, for instance, can serve as explicit communicative tools: they are public, costly, and often paired with formal statements that articulate the sender's demands and conditions for their removal. By contrast, clandestine cyber operations may quietly impose friction or degradation on the target's capabilities, while sanctions or diplomatic communication provide the narrative frame through which the intent is conveyed. In this way, cyber and non-cyber instruments can operate in tandem - one acting in the shadows to alter the material or cognitive conditions, the other performing the visible communicative function necessary for coercion or deterrence.

Model

Set up

The model outlined in this paper represents the use of coercive sanctions and cyber operations (i.e. clandestine distortionary attacks) as a game in which one state - state A - chooses whether to launch a cyber attack and/or whether to implement sanctions on another state - state B - in order to persuade them to cease efforts working towards accomplishing some program objective. This objective could be anything for which state B's success is uncertain to both them and state A, and state A does not want state B to accomplish it; the main example from this paper is Iran's nuclear program, but the model could be applied more broadly to any situations that share similar features.

State A is either highly capable at launching cyber attacks or relatively less capable at launching cyber attacks, parameterized by a cost of launching an attack $c_H < c_L$. State B does not observe state A's type; their prior beliefs are that low types occur with probability p and high types occur with probability $1-p$. Both types also pay a cost c_S for imposing sanctions on state B, which are threatened in advance and imposed only when state B does not acquiesce to the threat.

State B has program capabilities (e.g. nuclear capabiltiies) b which are continuous distributed $b \sim f[0,1]$, which represent the probability of their program succeeding (i.e. $Pr(success) = b$). However, they do not observe these capabilities directly; instead, they observe a signal θ that is a function of both their actual capabilities and a potential cyber attack that can partially disrupt their signal. Specifically, the signal takes the functional form:

$$\theta = \frac{b}{\gamma \mathbb{1}_K + 1}$$

Where $\mathbb{1}_K \in \{0,1\}$ is an indicator function for when an attack K occurs, and $\gamma \in \mathbb{R}^+$ is the degree to which an attack degrades the signal; higher levels of γ degrade the signal more. Thus, for any b, $\theta < b$ when a cyber attack occurs, and $\theta = b$ when a cyber attack does not occur. In this model, the cyber attack does not do any direct physical damage; it is purely a distortionary attack that operates through informational channels, in order to focus our attention and better understand this particular mechanism and its implications.

State B does not directly observe whether a cyber attack was launched. Key in this model is that in order for this signal-disrupting cyber attack to have any effect on state B's assessment of their probability of success, there has to be uncertainty as to whether a cyber attack is launched; this requires both that the attack itself is unobservable, and that there be some states of the world in which a cyber attack is not launched.

This leads to the following expected utility calculations for state B of rejecting demands from state A or acquiescing to them, where $V \in \mathbb{R}^+$ is the value they receive from achieving their objective, $S \in \mathbb{R}^+$ is the cost of sanctions, and $\mathbb{1}_S \in \{0,1\}$ is an indicator function for when sanctions are imposed:

$$EU(Reject) = bV - S\mathbb{1}_S$$

$$EU(Acquiesce) = 0$$

And the following timeline for the model:

1. Nature determines whether state A's cyber capabilities are high or low.

2. State A observes this, and determines whether or not to launch a cyber attack and whether or not to threaten sanctions.

3. State B observes State A's choice of whether to threaten sanctions, and a signal θ, and then chooses whether or not to Reject the demands of State A or Acquiesce to them.

4. If State B Rejects, sanctions are imposed if they were threatened, creating costs for both state A and state B. The outcome of the program is realized, leading to a payoff of V for State B if they succeed. State A gets a payoff of $X \in \mathbb{R}^+$ if the program objective is not achieved, either because State B abandons pursuing it or if State B tries and fails.

Analysis

Of primary interest in this paper are the equilibria in which cyber attacks distort state B's beliefs about their probability of success, leading to a higher probability of acquiescing to state B's demands in the face of sanctions. In other words: we are interested in exploring the conditions under which cyber attacks can be complementary to economic coercion, where cyber attacks in this model are a clandestine way of distorting another's state's perceptions of their own likelihood of success in some goal.

To explore these conditions, we will characterize the set of Perfect Bayesian Nash equilibria that (1) pool across state A's types in the decision to threaten sanctions; (2) separate across state A's types in the decision to employ a cyber attack. If separation occurred in the threatened use of sanctions, this would perfectly reveal state A's type to state B; then, even if types differ in their use of cyber attacks, state B would always know when a cyber attack was being employed and would be able to correct for it in their ex post evaluation of their probability of success. Similarly, if pooling occurs either (1) on no cyber attacks; (2) on launching cyber attacks, then state B knows with certainty whether or not a cyber attack has occurred and can correct for it.

Thus, beginning with the conjecture that both high and low types of state A threaten sanctions but only high types launch cyber attacks, we can consider state B's posterior beliefs about their probability of success conditional on observing their signal θ. If $\theta > \frac{1}{\gamma+1}$, then they know with certainty that state A must be a low type; this is because the highest possible signal they can receive if a cyber attack is launched is what they would receive if $b = 1$, so any signal higher than this must imply that a cyber attack has not been launched, which under the conjecture we are considering only happens when state A is a low type.

When $\theta \leq \frac{1}{\gamma+1}$, state B knows there is a probability p chance they are facing a low type and their signal is undistorted, and a probability $1-p$ chance they are facing a high type and their signal is being distorted. This leads to the following posterior belief about their probability of success:

$$Pr(Success|\theta) = p\theta + (1-p)(\gamma+1)\theta$$

Which allows us to characterize state B's expected utility from Rejecting state A's demands upon observing θ as:

$$EU(Reject) = V(p\theta + (1-p)(\gamma+1)\theta) - S$$

As they receive zero utility with certainty if they Acquiesce to demands, they thus Reject whenever the above expression is positive. Setting this equal to zero and rearranging allows us to determine their cutoff signal for Acquiescing as:

$$\bar{\theta} = \frac{S}{V(p + (1-p)(\gamma+1))}$$

We will assume $\bar{\theta} \leq \frac{1}{\gamma+1}$ for simplicity; if it exceeds this fraction, then a cyber attack by state A ensures that all types of state B Acquiesce.

We now consider the decision of state A of whether to impose sanctions or not. For the current conjecture to hold, we want it to be incentive compatible for both types of state A to threaten sanctions; we characterize the expected utility to low types from threatening sanctions under the current conjecture as the following:

$$EU(sanctions) = \underbrace{\int_0^{\bar{\theta}} Xf(b)db}_{\alpha} + \underbrace{\int_{\bar{\theta}}^1 (1-b)Xf(b)db}_{\beta} - \underbrace{\int_{\bar{\theta}}^1 c_S f(b)db}_{\delta}$$

The first part of this expression (α) is the payoff to state A obtained from state B acquiescing; it therefore integrates over all b such that state B acquiesces. The second part of this expression (β) is the payoff to state A obtained from state B pursuing their program but failing to achieve it. The final part (δ) is the cost of sanctions, which are only realized when state B rejects state A's demands, leading to sanctions imposition.

The expected utility of refraining from threatening sanctions is simply the expected utility obtained from hoping the program fails to materialize on its own, so:

$$EU(\text{no sanctions}) = \int_0^1 (1-b)Xf(b)db$$

Therefore, sanctions are threatened by low types in this conjecture whenever:

$$EU(sanctions) - EU(\text{no sanctions}) = \int_0^{\bar{\theta}} Xf(b)db + \int_{\bar{\theta}}^1 (1-b)Xf(b)db \int_{\bar{\theta}}^1 c_S f(b)db - \int_0^1 (1-b)Xf(b)db > 0$$

Which we can rewrite as the following, decomposing the integrals to allow for a clearer comparison:

$$\int_0^{\bar{\theta}} bXf(b)db+\int_0^{\bar{\theta}}(1-b)Xf(b)db+\int_{\bar{\theta}}^1(1-b)Xf(B)-\int_{\bar{\theta}}^1 c_S f(b)db-\int_0^{\bar{\theta}}(1-b)Xf(b)-\int_{\bar{\theta}}^1(1-b)Xf(b)db>0$$

Which simplifies to the following:

$$\int_0^{\bar{\theta}} bXf(b)db-\int_{\bar{\theta}}^1 c_S f(b)db>0$$

The first part of this integral is the added value to low types of threatening sanctions; for any given probability of success b, the expected payoff to state A of persuading state B to acquiesce is that probability of success times the payoff of avoiding the program X. The first part therefore integrates over the range of b that will be induced to Acquiesce by the imposition of sanctions. The second part integrates over the possible values of b in which sanctions have to be imposed after threatening, thus representing the marginal cost of threatening sanctions. Low type state As thus threaten sanctions when this expression for the net impact on utility is positive.

Under the current conjecture, we have that high types launch cyber attacks while low types do not; since $\theta=\frac{b}{\gamma+1}$, this has the effect of changing the cutoff value for b which leads to acquiescing to $\bar{\theta}(1+\gamma)$, since this is the value for b that will lead state B to observe $\bar{\theta}$ under conditions of a cyber attack. Since $\bar{\theta}(1+\gamma)>\bar{\theta}$ and nothing else about the sanctions analysis above changes, we have derived the following Lemma.

Lemma 1. *Under a conjecture where high and low type State As separate in the use of cyberattacks, both types pool on threatening sanctions iff:*

$$\int_0^{\bar{\theta}} bXf(b)db-\int_{\bar{\theta}}^1 c_S f(b)db>0$$

Proof follows from preceding discussion. We now turn to establishing the conditions under which, conditional on sanctions being threatened by both types, it is incentive compatible for high types, and only high types, to launch cyber attacks. Performing similar comparisons of integrals as above, we derive the following proposition.

Proposition 1. *A Perfect Bayesian Nash Equilibrium exists where both high and low types of State A threaten sanctions but only high types launch cyber attacks, and State B Acquiesces when $\theta \leq \bar{\theta}$ and Rejects otherwise, when the conditions of Lemma 1 are met in addition to the following condition:*

$$c_L > \int_{\bar{\theta}}^{\bar{\theta}(1+\gamma)} bXf(b)db + \int_{\bar{\theta}}^{\bar{\theta}(1+\gamma)} c_S f(b)db > c_H$$

Proof. For high types, we compare the expected utility conditional on launching a cyber attack and threatening sanctions with the expected utility of deviating to not launching a cyber attack:

$$EU(\text{cyber attack}) = \int_0^{\bar{\theta}(1+\gamma)} Xf(b)db + \int_{\bar{\theta}(1+\gamma)}^1 [(1-b)(X) - c_S]f(b)db - c_H$$

$$EU(\text{no cyber attack}) = \int_0^{\bar{\theta}} Xf(b)db + \int_{\bar{\theta}}^1 [(1-b)(X) - c_S]f(b)db$$

Decomposing integrals and subtracting the second expression from the first gives that launching a cyber attack is a best response whenever:

$$\int_{\bar{\theta}}^{\bar{\theta}(1+\gamma)} bXf(b)db + \int_{\bar{\theta}}^{\bar{\theta}(1+\gamma)} c_S f(b)db > c_H$$

For low types of State A the analysis is similar, except that they pay a cost $c_L > c_H$ to launch a cyber attack. So to ensure separation, we ensure that the value of the left hand side of this expression falls between c_H and c_L. □

The condition required for a PBNE is an intuitive one: the first integral computes the added expected utility that a cyber attack generates from persuading state B to Acquiesce in cases where they would have otherwise succeeded in their program goals, while the second integral is the added utility that a cyber attack generates by reducing the number of instances where sanctions actually need to be imposed after being threatened (which also harms state A with cost c_S). Locating this added value between c_H and c_L ensures separation.

At this point, we can consider a few different notions of "effectiveness" of cyber attacks and evaluate whether they hold in this equilibrium. First, we can consider whether in the context of this equilibrium sanctions are more effective in achieving their goals when cyber attacks are launched than when they are not; the answer to this is clearly yes, since $\bar{\theta}(1+\gamma) > \bar{\theta}$, and all draws of b between those two values lead state B to Acquiesce under cyber attacks but to Reject when a cyber attack is not launched.

The key intuition here is that State B is uncertain about state A's cyber capabilities, and thus cannot fully discern whether their signal about the likelihood of some program's success is the result of objective facts of the situation or the distortions created by an adversary. This allows the cyber attack to increase the probability that state B gives up their program in the face of an attack.

However, an interesting feature of this equilibrium is that the cutoff in the absence of a cyber attack is *higher* than it would have been if cyber attacks were not possible, or if they were never launched in equilibrium. If cyber attacks are never launched, then $\theta = b$, and the cutoff value for Acquiescing under threat of sanctions is:

$$\bar{b} = \frac{S}{V}$$

Which if we compare to $\bar{\theta}$:

$$\bar{\theta} = \frac{S}{V}\frac{1}{p+(1-p)(\gamma+1)}$$

It is clear that $\bar{b} < \bar{\theta}$, an implication of which is that low types of state A are always worse off if cyber attacks are possible and known to be employed by some types of state A. This makes sense; if state B is adjusting their estimates of success for the possibility of a cyber attack, then they will assume they have a better chance of success than their signal would otherwise indicate. A policy implication of this is that states should be very cautious about persuading other states that they may have capabilities that they do **not** actually have.

Of key interest, however, is whether or not high types do better at compelling Acquiescence in a separating equilibrium in which they launch cyber attacks than they would in the world in which cyber attacks did not exist (or the payoff-equivalent world of such attacks existing but with pooling on not using them). This leads to the following proposition:

Proposition 2. *There is a higher likelihood that State B Acquiesces when cyber attacks occur in the separating equilibrium of Proposition 1 than when cyber attacks are not possible.*

Proof. Here we simply need to compare $\bar{b}$ with $\bar{\theta}(1+\gamma)$. With some rearranging we get:

$$\bar{\theta}(1+\gamma) = \frac{S}{V\left(\frac{p}{\gamma+1}+(1-p)\right)}$$

Which is clearly greater than $\bar{b} = \frac{S}{V}$ since $\frac{p}{\gamma+1}+(1-p) < 1$. □

This proposition provides the key policy implication of the paper; namely, that cyber attacks can be a way of making sanctions more effective. Another proposition follows quickly:

Proposition 3. *Increases in p (the prior belief that State A has low cyber capabilities) increase the effectiveness of cyber attacks, where "effectiveness" is defined as the difference in probability of state B Acquiescing under the separating equilibrium of Proposition 1 versus a pooling equilibrium on no cyber attacks. Conversely, as p approaches zero, the effectiveness of cyber attacks converges to zero, but it remains an equilibrium for State A to keep using them; for sufficiently low p, State A is worse-off than if they did not have cyber capabilities and could credibly demonstrate that.*

Proof. We have established in the previous proposition that the separating equilibrium leads to more Acquiescence, but by how much? To determine the impact of p, we differentiate $\bar{\theta}(1+\gamma)$ with respect to p:

$$\frac{\partial\bar{\theta}(1+\gamma)}{\partial p} = \frac{S}{V}\left(\frac{p}{1+\gamma}+(1-p)\right)^{-2}\left(\frac{1}{1+\gamma}-1\right)(-1)$$

Since $\frac{1}{1+\gamma}-1<0$, this whole expression is positive, thus establishing the cutoff is increasing in p, resulting in more state Bs Acquiescing.

For the second part of the proposition, we can start by simply noting that the cutoff signal converges to $\bar{b}$ as p approaches zero, i.e.:

$$\lim_{p\to 0}\frac{S}{V\left(\frac{p}{\gamma+1}+(1-p)\right)} = \frac{S}{V} = \bar{b}$$

Clearly if $\bar{\theta}(1+\gamma) = \bar{b}$, State A obtains no benefit from having cyber capabilities and using them relative to not having those capabilities and having that be common knowledge, but they are paying the additional cost c_H from launching these attacks; this establishes that there exists p sufficiently low where State A is worse off because they have cyber capabilities. We can also characterize the condition under which State A is made worse off by having cyber capabilities:

$$\int_{\bar{b}}^{\bar{\theta}(1+\gamma)} bXf(b)db + \int_{\bar{b}}^{\bar{\theta}(1+\gamma)} c_S f(b)db < c_H$$

□

The intuition of this proposition is straightforward; as State B becomes more convinced that State A does **not** have cyber capabilities, the use of those cyber capabilities becomes more effective, since State A incorporates the possibility of a cyber attack into their calculations less. This provides an interesting policy implication; states developing the capabilities to clandestinely attack another state have an interest in keeping those capabilities secret! This may counterbalance the incentive to make such capabilities known for the purposes of deterrence.[3]

On the other hand, if State B becomes sufficiently convinced that State A **does** have cyber capabilities, State A is faced with a conundrum, where they are still incentivized to launch cyber attacks given that otherwise State B will be especially emboldened to pursue their project, but they are actually *worse off* than if they could credibly and publicly give up their capabilities in cyber. This of course will be difficult to do credibly, since a better outcome for them is if they can persuade State B that they do not have such capabilities while retaining them.[4] This result holds even without taking into account any fixed costs to developing such capabilities in the first place.

Summary of the Intuition

What have we learned from the model? Cyber attacks - interpreted in this context as **clandestine** attacks that distort an adversary's perceived probability of their own success in

[3] See, for instance, Kostyuk (2021) on public cyberinstitutions as a means of cyber deterrence.

[4] This result is also replicated in the pooling equilibrium in which both high and low types launch cyber attacks.

achieving some goal - can make coercive economic sanctions more likely to achieve their goals. Moreover, states have an interest in keeping their capabilites of launching cyber attacks **secret**; the revelation of their capabilities makes them less effective, since adversaries will start to assume that they are being attacked, and adjust their self-evaluations accordingly. In the limit, this actually makes states **worse off** than if they did not have these capabilities in the first place, or could credibly abandon them. These implications hold for any situations that share these features, including non-cyber clandestine attacks that perform a similar signal-distorting role.

Stuxnet: Clandestine Cyber Operations in Iran

Iran's nuclear program, particularly its uranium enrichment activities at facilities like Natanz and Fordow, has long been a focal point of international tension. The enrichment process, involving thousands of delicate centrifuges, is technically complex and vulnerable to disruption. Western powers, especially the United States and Israel, viewed Iran's progress toward potential weapons capability as a threat to regional and global security. By the mid-2000s, diplomatic efforts and economic sanctions had failed to halt Iran's enrichment activities, leading policymakers to consider more drastic measures.

Evidence suggests that the Obama Administration decided to accelerate the cyber attacks against Iran that initially began during the Bush Administration under the code name Olympic Games, in order to slow down the development of the Iranian nuclear program (Sanger, 2012). At the time, Stuxnet was considered alongside other possibilities, such as economic sanctions and threats of military force. Given that the U.S.'s European allies were divided on the issue of punitive sanctions on an already weak Iranian economy, especially in the aftermath of the 2003 Iraq war, the United States and Israel seemingly chose cyber tools as the most viable course of action. Launched in 2006 during the Bush Administration (Kaplan, 2016), then-Vice President Cheney advocated for launching air strikes against the

Natanz reactor. Bob Gates, the Defense Secretary, in turn “persuaded Bush that going to war against a third Muslim country, while the two in Afghanistan and Iraq are still raging, would be bad for national security.” To wit, Stuxnet seemed like the Goldilocks option as Bush was looking for a ‘third option’ - “something in between air strikes and doing nothing” (Kaplan, 2016).

According to reports, President Obama similarly believed that the United States had no other choice than stopping Iran from developing nuclear capabilities (especially after Iran had previously claimed that it stopped pursuing nuclear weapons in 2003). “If Olympic Games failed... [Obama told aides], there would be no time for sanctions and diplomacy with Iran to work. Israel could carry out a conventional military attack, prompting a conflict that could spread throughout the region” (Sanger, 2012).

A highly sophisticated computer worm, dubbed Stuxnet by the computer-savvy civilians who first brought it to public attention, significantly damaged Iran’s uranium enrichment infrastructure in Natanz. Stuxnet was a cyber attack that targeted the supervisory control and acquisition (SCADA) systems, overcame the air-gap protections of the Iranian nuclear facility, and disabled the centrifuges engaged in the purification of uranium. It was customized to manipulate the programmable logic controllers (PLCs), and alter the rotational speed of IR-1 centrifuges (Kaplan, 2016; Peterson, 2013; Lindsay, 2013; Zetter, 2015*a*). Highly targeted and highly sophisticated, Stuxnet utilized four zero days - vulnerabilities that were previously unknown to software developers. While affecting the underground facility with an air strike would have been possible, such course of action carried more risk (Lindsay, 2013). The immediate effect of Stuxnet was the physical destruction of centrifuges and a setback to Iran’s enrichment timeline. Indeed, this was the first known international cyber attack to cause physical damage to nuclear infrastructure.

However, the psychological and strategic ramifications were even more profound. Stuxnet’s clandestine nature meant that, for months, Iranian engineers struggled to diagnose the cause of repeated equipment failures. The worm's ability to falsify operational data further confounded efforts to identify the source of the problem.

While the Stuxnet attack is well-documented elsewhere (Kaplan, 2016; Peterson, 2013; Lindsay, 2013; Zetter, 2015*a*), what is less explored is the *strategic* effect Stuxnet had beyond disabling the Iranian enrichment program for the time period of less than a year. In accordance with our theoretical predictions, the most significant effect of Stuxnet was the uncertainty it injected into Iran's strategic calculus. Prior to the attack, Iran's leadership likely believed that, given enough time and resources, they could overcome technical hurdles and achieve their nuclear objectives. Stuxnet undermined this confidence by making it seem as though their scientists simply might be unable to deliver regardless of the level of investment. Indeed, this was the intent of the attack - as one participant put it under conditions of anonymity:

> "The idea was to string it out as long as possible. If you had wholesale destruction right away, then they generally can figure out what happened, and it doesn't look like incompetence.' '(Nakashima and Warrick, 2012)

Through cyber action, the United States and Israel, who are alleged to be behind the Stuxnet virus, managed to, in the time of peace, "manipulate the strategic environment to [their] advantage while at the same time degrading the ability of an adversary [Iran] to comprehend that same environment" (Sheldon, 2011). The success of the U.S. and Israel in disabling the Iranian nuclear program can largely be ascribed to the dependence of Iranian infrastructure on cyberspace. Thus, Stuxnet created enough uncertainty that even if Iran wanted to continue to develop the program and not come to the negotiating table, Tehran was uncertain that they would ever be successful in their efforts. Indeed, as our formal model suggests, if an actor can create enough uncertainty about the likelihood of victory, its expected utility decreases.

In hindsight, the Stuxnet virus may have ultimately been successful in its aims, despite its relatively early and unanticipated discovery. Stuxnet was first identified in mid-September 2010 (although the operation itself allegedly began in 2007), with the full scope of the operation revealed over the course of 2011-2012. These early years of the program proved crucial to bringing Iran to the bargaining table; prior to 2007, Iran appeared to have little interest in negotiating over restrictions to their nuclear program, and indeed fell out of compliance with the Treaty on the Non-Proliferation of Nuclear Weapons that they had signed in 1970 (Lindsay, 2025, p.128). While the details of secret talks between the US and Iran in the years that followed remain highly concealed, as Obama era officials were apparently concerned at the time about angering their Israeli allies (Bergman and Mazzetti, 2019), it seems as though Iran became more receptive to negotiating in the years that followed the beginning of the Stuxnet operation, engaging in secret talks with Obama administration officials over the potential contours of such a deal (Calabresi, 2009). At least by late 2010, these secret talks were maturing into a backchannel process that, once underway, would lead to intensified negotiations in the years that followed (Lindsay, 2025, p.129)

By 2011, Iran started negotiating with P5+1. In 2013, there were the first signs that the two sides might be able to reach a comprehensive agreement. Iran eventually signed the Joint Comprehensive Plan of Action (JCPOA) in 2015 agreeing to almost fully eliminate its stockpiles of medium and low-enriched uranium, and diminish its nuclear infrastructure. Further, while the deal was in effect, Iran was primed to reenter the international community with the ultimate lifting of the oil embargo and sanctions relief. Until the Trump Administration's unilateral withdrawal from the agreement, it had been undeniably effective and successful.

Would these talks have ever even begun if Stuxnet had not undermined the confidence that Iranian scientists could develop a nuclear program on their own? While it is impossible to know the counterfactual, and difficult to observe all of the early events that ultimately led to the signing of the JCPOA, we can at least observe that at some point over the period leading up to 2010, Iran appears to have become more willing to negotiate over a sanctions-relief-for-concessions deal, and we can see that these negotiations ultimately produced exactly this kind of deal. While Stuxnet was ultimately revealed during the course of these negotiations, by this point the negotiations had their own momentum, having overcome Iranian reluctance to consider any concessions on principle.Indeed, as the New York Times reports:

> "The secrecy around the talks remains a freighted subject among many former Obama officials, one that few are willing to discuss on the record. Some believed that the Obama-Netanyahu relationship had grown so toxic that the Israeli prime minister couldn't be trusted. And, they argue, the strategy worked: Talks stayed quiet long enough for them to mature into serious negotiations and, ultimately, the Joint Comprehensive Plan of Action." (Bergman and Mazzetti, 2019)

Stuxnet's legacy extended far beyond the immediate damage it inflicted at Natanz. The operation demonstrated the potential of cyber capabilities to achieve strategic objectives without resorting to traditional military force.

For Iran, the experience of Stuxnet was a lesson in the vulnerabilities inherent in complex, interconnected systems. The regime began to invest further in cybersecurity, and reconsidered its technical and operational practices. For other states, Stuxnet served as both a warning and a template, spurring investments in both offensive and defensive cyber capabilities.

Expectations of Distortions in Iran and North Korea

With the details of Stuxnet now in the public record, the capabilities of the United States and Israel to launch similar attacks in the future were clearly revealed to the world. Our model predicts that the revelation of capabilities (the model would represent this with p close to zero) should result in two clear outcomes: (1) adversaries will begin to incorporate the likelihood of sabotage into their assessments of their probability of success in various programs that the United States/Israel is pressuring them to abandon; (2) the effectiveness of clandestine distortionary attacks is consequently likely to decline.

One place we see evidence of these predictions is in Iran itself. In the years since Stuxnet's revelation, Iran has responded to several facially random accidents with accusations of sabotage. In 2020, after a fire broke out at Natanz, Iran immediately accused Israel of responsibility, prompting the response from Israeli Defense Minister Benny Gantz that "Not every event that happens in Iran is necessarily related to us." (Fassihi, Pérez-Peña and Bergman, 2020) In 2021, after Natanz faced a blackout, Malek Shariati Niasar, the spokesman for the Iranian Parliament's energy committee, took to Twitter to note that the outage was "very suspicious", and raised the possibility of "sabotage and infiltration." (Bergman, Gladstone and Fassihi, 2021) Similarly, Mohsen Rezaei, a former commander-in-chief of the Islamic Revolutionary Guard Corps tweeted "Could the reoccurrence of a fire at the Natanz nuclear facilities, in less than one year following the previous explosion, be a sign of the seriousness of the infiltration phenomenon?" (Motamedi, 2021) Passing an attack off as incidental appears to have become more difficult as Iran has become hyper-vigilant about clandestine sabotage operations.

Moreover, while Stuxnet's clandestine distortions may have been instrumental to bringing Iran to the table for nuclear negotiations, the 2021 possible attacks had seemingly *no impact* on negotiations that were ongoing between the US and Iran about a potential new deal. Iranian Foreign Minister Mohammad Javad Zarif was quoted as saying to Iranian lawmakers:

> "Now they think they will achieve their goal. But the Zionists will get their answer in more nuclear advancements. If they think our hand in the negotiations has been weakened, actually this cowardly act will strengthen our position in the talks. Other parties to the talks must know that if they faced enrichment facilities that used first-generation machines, now Natanz can be filled with advanced centrifuges that have several times the enrichment capacity." (Motamedi, 2021)

A far cry from the recriminations and self-blame that were reported in the face of Stuxnet, in a way that accords well with the predictions of this paper's model. Indeed, Biden era negotiations with Iran ultimately stalled and did not produce a return to a JCPOA-like agreement, in part due to Iranian unwillingness to make early concessions; instead, Iran escalated their enrichment during negotiations (Kingston, 2024).

Another useful test case is the unverified but possible "left-of-launch" cyber operation against North Korea (**?**, p.268) – dubbed by some as a "Stuxnet for North Korea" (Zetter, 2015*b*). Left-of-launch is contrasted with "right-of-launch" operations that target and defend against missiles after they've been launched; instead, left-of-launch operations target the supply chains and operational systems that govern a missile's success before liftoff or during the first seconds of flight (Broad and Sanger, 2017).

In the wake of Stuxnet, there has been discussion, but mixed evidence, of a potential attempt to use cyber operations to clandestinely sabotage North Korean missiles in a left-of-launch fashion (Sanger and Broad, 2017; Lewis, 2017) Moreover, in the early stages of the North Korean nuclear missile program, their missiles were decidedly *not* effective; after the first eight tests, only one succeeded (Sanger and Broad, 2017). Like with Stuxnet, we have a similar pattern of failure, where the rates of failure are plausibly the result of human or random error, but are higher than one might otherwise expect.

If such cyber attacks had been launched by the US against North Korea – and intelligence officials and experts differ strongly on the likelihood that such a program was seriously attempted, such that the existence of this program remains highly uncertain – then the model would predict that North Korea's knowledge that such an attack *might* be launched would limit its effectiveness, as they took into account this probability.

As the New York Times reports, "It was after the last failure that the North Korean leader, Kim Jong-un, was reported to have ordered an investigation into whether the United States was sabotaging his country's test flights, searching for spies in his system." (Sanger and Broad, 2017) Furthermore, in stark contrast with Iran pre-Stuxnet's discovery, where they fired engineers in response to the high failure rate at Natanz, Kim Jong-Un openly celebrated the nuclear missile scientists, building luxury apartment towers for them, holding galas and rallies in their honor, sharing cigarettes with several of them, and in at least one case giving a scientist a piggy-back ride (Sang-Hun et al., 2017).

Moreover, unlike Iran, which appears to have been made more willing to bargain at least in part due to the clandestine distortions of Stuxnet, North Korea continued on with their missile program unabated, and now appears to have largely succeeded in mass producing intermediate-range ballistic missiles (Wertz and Lewis, 2024). This also accords well with the predictions of the model, in which the revelation of capacities leads potential targets to expect these unobserved distortions, in a way that mitigates such distortions' ability to be effective.

Conclusion

In this paper, we have explored the strategic implications of attacks that take the form of *clandestine distortions*; a kind of attack that has become much more easy to implement with the advancement of cyber technology. We have demonstrated that such attacks, despite the fact that they are both unattributed and unobserved, can still achieve coercion by complementing overt forms of coercion, degrading the benefit side of a cost-benefit calculus posed by a coercive threat. However, we have also demonstrated that the impact of such attacks depends crucially on an adversary's uncertainty that such attacks could be launched by the distortion-sender; as they become more confident that such an attack *can* be sent by a sender, they start to assume that attacks are being sent with some probability, and thus adjust their expectations of success in their project to be higher than whatever signal they receive. This can ultimately undermine the effectiveness of such attacks, making low-capabilities states worse-off than they would be if the technology simply did not exist, and can even be welfare reducing for the state with high cyber capabilities as awareness grows of their capabilities. An equilibrium in which distortions are endemic but ineffective and costly is possible in both separating and pooling equilibria of the model.

Nonetheless, despite these constraints, the model also helps us see why such attacks can sometimes be very useful, particularly in cases where the program in question is very valuable to the target. Pakistani Prime Minister Zulfiqar Ali Bhutto once famously remarked that "Pakistan will eat grass or leaves" if that was what was necessary to obtain nuclear weapons (Anderson and Khan, 1998). Attempting to get such an adversary to abandon a program by escalating costs alone is unlikely to be effective, but the chances may be greater if the expected benefit of pursuing a program can be sufficiently degraded by generating enough uncertainty about whether achieving such a program is even possible.

As an illustration, consider the 2025 military conflict between Israel and Iran, which became known as the Twelve-Day War. It began when Israel launched a series of surprise attacks against Iran's nuclear program, targeting facilities, military assets, and personnel, and escalated further when the United States joined the conflict by conducting airstrikes on the key Iranian nuclear facilities at Fordow, Natanz, and Isfahan. Relying on satellite imagery and other information, the IAEA has confirmed significant damage to several sites - indeed, significantly more *physical* damage than was ever inflicted by the Stuxnet virus

(Bertrand, Lillis and Cohen, 2025). However, at least in the immediate wake of these attacks, it appears that Iran has been largely undeterred from the pursuit of their nuclear program, and has instead moved facilities further underground and redoubled their efforts to rebuild (Erlanger, 2025). Indeed, Iran also appears to have expressed little increased willingness to bargain over the program, with Foreign Minister Abbas Araghchi rejecting US terms for negotiations, stating "We will never negotiate our missile program, and no rational actor would disarm. We cannot stop uranium enrichment, and what cannot be achieved by war cannot be achieved through politics. We have no desire for direct talks with Washington" (Staff, 2025). It thus at least appears that these much more physically damaging attacks may ultimately prove to be less effective than Stuxnet's more subtle approach of generating uncertainty about the regime's ability to achieve a program at all.

The nature of the kind of attacks we characterize in this paper means that we may never be made aware of many of the most important examples, but we are especially confident that understanding these mechanisms will be important to understanding the future of strategic competition between states as cyberspace becomes an increasingly prominent domain in which such competition unfolds.

References

1947 National Security Act. 1947.

Allen, Susan Hannah. 2005. "The Determinants of Economic Sanctions Success and Failure." *International Interactions* 31(2):117–138.

Allen, Susan Hannah. 2008. "The Domestic Political Costs of Economic Sanctions." *Journal of Conflict Resolution* 52(6):916–944.

Anderson, John Ward and Kamran Khan. 1998. "Pakistani Politicians' Rallying Cry: 'Let Them Eat Grass'.".

Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzky. 2020. "Deterrence with Imperfect Attribution." *American Political Science Review* 114(4):1155–1178.

Bapat, Navin A. and T. Clifton Morgan. 2009. "Multilateral Versus Unilateral Sanctions Reconsidered: A Test Using New Data." *International Studies Quarterly* 53(4):1075–1094.

Bapat, Navin R, Bryan R Early, Julia Grauvogel and Katja B Kleinberg. 2024. "The Currency Constraint: Explaining the Selective Enforcement of US Financial Sanctions." *Foreign Policy Analysis* 20(4).

Bergman, Ronen and Mark Mazzetti. 2019. "The Secret History of the Push to Strike Iran.".

Bergman, Ronen, Rick Gladstone and Farnaz Fassihi. 2021. "Blackout Hits Iran Nuclear Site in What Appears to Be Israeli Sabotage.".

Bertrand, Natasha, Katie Bo Lillis and Zachary Cohen. 2025. "Early US intel assessment suggests strikes on Iran did not destroy nuclear sites, sources say." *CNN* .

Blainey, Geoffrey. 1988. *Causes of war*. Simon and Schuster.

Borghard, Erica D and Shawn W Lonergan. 2017. "The logic of coercion in cyberspace." *Security Studies* 26(3):452–481.

Broad, William J. and David E. Sanger. 2017. "U.S. Strategy to Hobble North Korea Was Hidden in Plain Sight.".

Brooks, Risa A. 2002. “Sanctions and Regime Type: What Works, and When?” *Security Studies* 11(4):1–50.

Buchanan, Ben. 2020. *The hacker and the state: Cyber attacks and the new normal of geopolitics*. Harvard University Press.

Calabresi, Massimo. 2009. “Obama’s Secret Iran Talks: Setting the Stage for a Deal?”.

Carson, Austin. 2018. Secret Wars. In *Secret Wars*. Princeton University Press.

CCDCOE. 2007. “Operation Orchard/Outside the Box (2007).”.

Cormac, Rory, Calder Walton and Damien Van Puyvelde. 2022. “What constitutes successful covert action? Evaluating unacknowledged interventionism in foreign affairs.” *Review of International Studies* 48(1):111–128.

Cormac, Rory and Richard J Aldrich. 2018. “Grey is the new black: covert action and implausible deniability.” *International affairs* 94(3):477–494.

Davis, Jason Sanwalka. 2021. “War as an Internal Information Problem.” *Working Paper* .

Davis, Jason Sanwalka. 2025. “Targeted Sanctions and Redistribution.” *Working Paper* .

DeVine, Michael E. 2022. “Covert Action and Clandestine Activities of the Intelligence Community: Selected Definitions.”.

Di Lonardo, Livio and Scott A Tyson. 2024. “Constraining to deter.” *American Journal of Political Science* .

Drezner, Daniel W. 1999. *The Sanctions Paradox: Economic Statecraft and International Relations*. Cambridge University Press.

Egloff, Florian J and Max Smeets. 2021. “Publicly attributing cyber attacks: a framework.” *Journal of Strategic Studies* pp. 1–32.

Erlanger, Steven. 2025. “The Dangerous Stalemate Over Iran’s Nuclear Program.”.

Falliere, Nicholas, Liam O’Murchu and Eric Chien. 2011. “W32.Stuxnet Dossier.”.

Fassihi, Farnaz, Richard Pérez-Peña and Ronen Bergman. 2020. “Iran Admits Serious Damage to Natanz Nuclear Site, Setting Back Program.”.

Fearon, James D. 1995. "Rationalist explanations for war." *International organization* 49(3):379–414.

Fischerkeller, Michael P, Emily O Goldman and Richard J Harknett. 2022. *Cyber persistence theory: Redefining national security in cyberspace*. Oxford University Press.

Gartzke, Erik. 1999. "War is in the Error Term." *International Organization* 53(3):567–587.

Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.

Gibson, William. 1984. *Neuromancer*. Springer.

Gross, Michael L, Daphna Canetti and Dana R Vashdi. 2017. "Cyberterrorism: its effects on psychological well-being, public confidence and political attitudes." *Journal of Cybersecurity* 3(1):49–58.

Harknett, Richard J and Max Smeets. 2022. "Cyber campaigns and strategic outcomes." *Journal of Strategic Studies* 45(4):534–567.

Jardine, Eric, Nathaniel Porter and Ryan Shandler. 2024. "Cyberattacks and public opinion–The effect of uncertainty in guiding preferences." *Journal of Peace Research* 61(1):103–118.

Jervis, Robert. 2017. *Perception and misperception in international politics: New edition*. Princeton University Press.

Joseph, Michael. 2025. "Unattributable Coercion." *Working Paper* .

Joseph, Michael F. 2023. "Do Different Coercive Strategies Help or Hurt Deterrence?" *International Studies Quarterly* 67(2).

Joseph, Michael F and Michael Poznansky. 2018. "Media technology, covert action, and the politics of exposure." *Journal of Peace Research* 55(3):320–335.

Jun, Jenny. 2021. "Coercion in Cyberspace: A Formal Model of Extortion via Encryption." *Working Paper* .

Kaplan, Fred. 2016. *Dark territory: The secret history of cyber war*. Simon and Schuster.

Kavakli, Kerim Can, J. Tyson Chatagnier and Emre Hatipoğlu. 2020. “The Power to Hurt and the Effectiveness of International Sanctions.” *The Journal of Politics* 82(3):879–894.

Kelton, Mark. 2025. “Here’s How Russia’s Covert War Could Undermine its Own Goals.”.

Kingston, Shannon K. 2024. “Biden administration throws cold water on prospect of renewed Iran nuclear talks.”.

Kostyuk, Nadiya. 2021. “Deterrence in the Cyber Realm: Public versus Private Cyber Capacity.” *International Studies Quarterly* 65(4):1151–1162.

Kostyuk, Nadiya and Yuri M Zhukov. 2019. “Invisible digital front: Can cyber attacks shape battlefield events?” *Journal of Conflict Resolution* 63(2):317–347.

Kustra, Tyler. 2023. “Economic sanctions as deterrents and constraints.” *Journal of Peace Research* 60(4):649–660.

Lektzian, David J. and Christopher M. Sprecher. 2007. “Sanctions, Signals, and Militarized Conflict.” *American Journal of Political Science* 51(2):415–431.

Lektzian, David and Mark Souva. 2007. “An Institutional Theory of Sanctions Onset and Success.” *Journal of Conflict Resolution* 51(6):848–871.

Lewis, Jeffrey. 2017. “Is the United States Really Blowing Up North Korea’s Missiles?” *Foreign Policy* .

Lindsay, Jon. 2020. “Military Organizations, Intelligence Operations, and Information Technology.” *Texas National Security Review* pp. 43–56.

Lindsay, Jon R. 2013. “Stuxnet and the limits of cyber warfare.” *Security Studies* 22(3):365–404.

Lindsay, Jon R. 2025. *Age of Deception: Cybersecurity as Secret Statecraft*. Cornell University Press.

Lindsay, Jon R and Erik Gartzke. 2018. “Coercion through cyberspace: the stability-instability paradox revisited.” *Coercion: The Power to Hurt in International Politics* pp. 179–203.

Maness, R., Valeriano B. Hedgecock K. Jensen B. M. Macias J. 2022. “Codebook for the Dyadic Cyber Incident and Campaign Dataset (DCID) version 2.0.”.

Marinov, Nikolay. 2005. “Do economic sanctions destabilize country leaders?” *American Journal of Political Science* 49(3):564–576.

Maschmeyer, Lennart. 2024. *Subversion: From covert operations to cyber conflict*. Oxford University Press.

McCormack, Daniel and Henry Pascoe. 2017. “Sanctions and Preventive War.” *The Journal of Conflict Resolution* 61(8):1711–1739.

McLean, Elena V and Taehee Whang. 2014. “Designing foreign policy: Voters, special interest groups, and economic sanctions.” *Journal of Peace Research* 51(5):589–602.

Morgan, T. Clifton and Valerie L. Schwebach. 1995. “Economic sanctions as an instrument of foreign policy: The role of domestic politics.” *International Interactions* 21(3):247–263.

Motamedi, Maziar. 2021. “Iran’s Zarif blames Israel for Natanz incident, vows revenge.”.

Nakashima, Ellen and Joby Warrick. 2012. “Stuxnet was work of U.S. and Israeli experts, officials say.”.

Nutt, Cullen G. 2025. “When Do Great Powers Employ Covert Action?” *Security Studies* 34(1):129–166.

of Staff, Joint Chiefs. 2021. “DOD Dictionary of Military and Associated Terms.”.

Oppenheimer, Harry. 2024. “How the process of discovering cyberattacks biases our understanding of cybersecurity.” *Journal of Peace Research* 61(1):103–118.

Peksen, Dursun. 2019. “Autocracies and Economic Sanctions: The Divergent Impact of Authoritarian Regime Type on Sanctions Success.” *Defence and Peace Economics* 30(3):253–268.

Peterson, Dale. 2013. “Offensive cyber weapons: construction, development, and employment.” *Journal of Strategic Studies* 36(1):120–124.

Poznansky, Michael and Evan Perkoski. 2018. “Rethinking Secrecy in Cyberspace: The

Politics of Voluntary Attribution." *Journal of Global Security Studies* 3(4):402–416.

Rid, Thomas and Ben Buchanan. 2015. "Attributing cyber attacks." *Journal of Strategic Studies* 38(1-2):4–37.

Riehle, Kevin P. 2020. "Russia's intelligence illegals program: an enduring asset." *Intelligence and National Security* 35(3):385–402.

Ross, Ron, Pillitteri Victoria Graubart Richard Bodeau Deborah and Rosalie McQuaid. 2021. "Developing Cyber-Resilient Systems: A Systems Security Engineering Approach.".

Rovner, Joshua, Rory Cormac and Lennart Maschmeyer. 2025. "Sand in the gears: Sabotage in world politics." *European Journal of International Security* pp. 1–20.

Sang-Hun, Choe, Motoko Rich, Natalie Reneau and Audrey Carlsen. 2017. "Rocket Men: The Team Building North Korea's Nuclear Missile.".

Sanger, David E. 2012. "Obama Order Sped Up Wave of Cyberattacks Against Iran.".

Sanger, David E. 2018. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. New York: Crown.

Sanger, David E. and William J. Broad. 2017. "Trump Inherits a Secret Cyberwar Against North Korean Missiles.".

Schelling, Thomas. 1960. *The Strategy of Conflict*. Harvard University Press.

Schrader, Stuart. 2018. "The Long Counterrevolution: United States-Latin America Security Cooperation.".

Schram, Peter. 2021. "Hassling: how states prevent a preventive war." *American journal of political science* 65(2):294–308.

Scott, Len. 2004. "Secret intelligence, covert action and clandestine diplomacy." *Intelligence & national security* 19(2):322–341.

Shandler, Ryan and Daphna Canetti. 2024. "Introduction: Cyber-conflict–Moving from speculation to investigation.".

Shandler, Ryan, Michael L Gross and Daphna Canetti. 2023. "Cyberattacks, psychological distress, and military escalation: An internal meta-analysis." *Journal of Global Security Studies* 8(1):ogac042.

Shandler, Ryan, Michael L Gross, Sophia Backhaus and Daphna Canetti. 2022. "Cyber terrorism and public support for retaliation–a multi-country survey experiment." *British Journal of Political Science* 52(2):850–868.

Sheldon, John B. 2011. "Deciphering cyberpower: Strategic purpose in peace and war." *Strategic Studies Quarterly* 5(2):95–112.

Spaniel, William. 2022. "Scientific intelligence, nuclear assistance, and bargaining." *Conflict Management and Peace Science* 39(4):447–469.

Staff, Toi. 2025. "Iran has 'no desire' for talks with US, 'cannot stop uranium enrichment,' says FM.".

Steiner, Peter. 1993. "On the internet, nobody knows you're a dog.".

Valeriano, Brandon, Benjamin M Jensen and Ryan C Maness. 2018. *Cyber strategy: The evolving character of power and coercion*. Oxford University Press.

Vićić, Jelena and Erik Gartzke. 2024. "Cyber-enabled influence operations as a 'center of gravity'in cyberconflict: The example of Russian foreign interference in the 2016 US federal election." *Journal of Peace Research* 61(1):10–27.

Vićić, Jelena and Richard Harknett. 2024. "Identification-imitation-amplification: understanding divisive influence campaigns through cyberspace." *Intelligence and National Security* 39(5):897–914.

Wertz, Daniel and Jeffrey Lewis. 2024. "North Korea's Ballistic Missile Program.".

Zetter, Kim. 2015*a*. *Countdown to zero day: Stuxnet and the launch of the world's first digital weapon*. Crown.

Zetter, Kim. 2015*b*. "The US Tried to Stuxnet North Korea's Nuclear Program.".