Faculty Society and Economics

HOCHSCHULE RHEIN-WAAL
Rhine-Waal University
of Applied Sciences

DISCUSSION PAPER

THE END OF INTERVENTIONS?

Simulating Cyberweapons as Deterrence Against Humanitarian Interventions

Christopher Gerritzen

No. 6

Discussion Papers in Behavioural Sciences and Economics

ISSN 2510-2729

Januar, 2020

Editors

Prof. Dr. Thomas Pitz, Hochschule Rhein-Waal, Faculty of Society and Economics, tel.: +49 2821 80673 337, email: thomas.pitz@hochschule-rhein-waal.de

Prof. Dr. Jörn Sickmann, Hochschule Rhein-Waal, Faculty of Society and Economics, tel.: +49 2821 80673 314, email: joern.sickmann@hochschule-rhein-waal.de

The discussion papers constitute work in progress. They are circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors.

The End of Interventions?

Simulating Cyberweapons as Deterrence Against Humanitarian Interventions

Christopher Gerritzen

Abstract

Cyberweapons are frequently discussed in politics, the media, and academic literature when looking at current and future conflict between states or state and non-state actors. It is often thought that cyberweapons – through their deployment in cyberwarfare or cyberterrorism, to name two examples – have the potential to drastically, perhaps even fundamentally, change international relations. A commonly found reasoning for this is the perception of cyberweapons as a tool that will alter the balance of power between states in a way that favors small states and non-state actors due to the vulnerability of high-tech states' IT systems. This paper will look at one particular scenario in which cyberweapons may have this effect and discuss whether small states can use cyberweapons as deterrence against humanitarian interventions. To answer this question, this paper will first look at the concept of deterrence and deterrence theory. Based on deterrence theory, a simple game of cybered deterrence will be developed taking into account a number of factors. First, the properties of cyberweapons themselves, based on five different cases of cyberattacks and theoretical examinations. Second, whether cyberweapons as deterrence can be credible and the escalation potential that may result from credibility. Third, the history of cyberweapons as a tool in international relations and why military interventions are sometimes stopped before completion. Lastly, the game will be both theoretically analyzed and simulated using reinforcement learning and belief learning. Based on these examinations, the paper will conclude that cyberweapons are at best an unreliable and risky deterrent against military interventions.

Keywords: military humanitarian interventions, cyberweapons, cyberdeterrence, cybered deterrence, cross-domain deterrence, simulations, reinforcement learning, belief learning

Acknowledgements

This paper was originally written as thesis under the title “Simulating Cyberweapons as Deterrence Against Humanitarian Interventions Using Game Theory and Learning Algorithms – Can Small States use Cyberweapons to Deter Interventions by Powerful, Networked States?” which was submitted in partial fulfillment of the requirements of the degree of Bachelor of Arts in International Relations at Hochschule Rhein-Waal in the winter semester of 2018/2019.

I am deeply grateful for the helpful support and suggestions of my supervisors during the writing of my thesis.

First, Prof. Dr. Thomas Pitz, in particular for supporting the development of the sections relating to the technical and theoretical foundations and simulations of the deterrence game. I also thank him for suggesting the publication of this thesis as discussion paper.

Second, Prof. Dr. Alexander Brand, especially for helping with the development of the research question itself to find a focus – humanitarian interventions – in the rather broad topic of cyberweapons as a means of deterrence.

I want to thank both very much for their support and for supervising my thesis.

Table of Contents

List of Abbreviations

BL	Belief Learning
CDD	Cross-Domain Deterrence
CDG	Cybered Deterrence Game
CNA	Computer Network Attack
CNE	Computer Network Exploitation
DDoS / DDOS	Distributed Denial of Service
DNS	Domain Name System
EWA	Experience Weighted Attraction
ICS	Industrial Control System
ICT	Information and Communications Technology
NATO	North Atlantic Treaty Organization
PDT	Perfect Deterrence Theory
PN	Powerful, Networked State (Player 1)
QRE	Quantal Response Equilibrium
RL	Reinforcement Learning
SCADA	Supervisory Control And Data Acquisition
SCDG	Sequential Cybered Deterrence Game
SM	Small State (Player 2)
US	United States
WFPM	Weighted Fictitious Play Model

List of Symbols

A_i^j	Attraction of player *i* to strategy *j*
c_{PN}	Variable cost of conflict of PN
c_{SM}	Variable cost of conflict of SM
I	Intervene (Strategy for PN/Player 1)
i	Player 1 (i = 1) or Player 2 (i = 2)
$I(.)$	Indicator function
j	Strategy
M	Maximum number of periods per simulation
N	Number of simulations
$N(t)$	Experience variable
P_i^j	Probability of player *i* to select strategy *j*
R	Retaliate (Strategy for PN/Player 2)
S	Surrender (Strategy for PN/Player 2)
$s_{-i}(t)$	Strategy of the opposing player
SD	Standard Deviation
$s_i(t)$	Strategy of player *i*
t	Specific period of the simulation
T (%)	Average percentage of simulations in the SCDG that were terminated after 10 periods, resulting in an automatic defeat of state SM
W	Withdraw (Strategy for PN/Player 1)
x_0	Payoff of PN for ‘Status Quo’
x_1	Payoff of PN for ‘SM Loses / PN Wins’ (strategy pair IS)
x_2	Maximum payoff of PN for ‘Conflict’ (strategy pair IR)
y_0	Payoff of SM for ‘Status Quo’ (strategy pairs WS and WR)

y_1	Payoff of SM for ‘SM Loses / PN Wins’ (strategy pair IS)
y_2	Maximum payoff of SM for ‘Conflict’
λ	Sensitivity to attractions
λ_{QRE}	Randomness or experience parameter in the QRE calculation
π	Payoff
ϕ	Recency parameter

List of Figures

List of Tables

1 Introduction

Cyberattacks, cyberwarfare, and cyberweapons[1] are only some terms that are increasingly featured in politics, the media, and academic literature when it comes to current and future conflict between states or state and non-state actors. Often, the terms are accompanied by fearful predictions of election hacking, catastrophic attacks rivaling Hollywood blockbusters, and generally the end of the world and international system as we know it. Of course, rarely does a week pass without reports of cyberattacks against sometimes high profile targets or newly discovered vulnerabilities in widely used software so those predictions do not appear too unrealistic at first glance. But do they hold up when investigated in more detail or are they in the end as questionable as the portrayal of hacking in the aforementioned Hollywood blockbusters?

One such prediction or idea about cyberattacks and cyberweapons is that they may be used as a powerful but relatively cheap deterrent, in particular by small states against powerful, networked states (Rustici, 2011; Gaycken and Martellini, 2013; Hughes and Colarik, 2016). It is based on the idea that "[t]he ability to reach out through cyberspace could negate the strategic advantage the United States has in the two oceans [because it] … puts United States' critical infrastructure at risk to attack from cyber-attack" (Rivera, 2012, p. 47). Because of this capability, "[c]yberweapons have the unique ability to change international relations in ways never seen before" (Rustici, 2011, p. 40) as it may change the balance of power between states significantly. Consequently, their use as a deterrent by small states may "… make the world a safer place for corrupt and abusive regimes" (Rustici, 2011, p. 38). This hypothesis will be analyzed in detail in this paper as it will attempt to answer whether small states can use cyberweapons to deter interventions by powerful, networked states. This will be done by applying game theory and simulating the game developed in this paper by using learning algorithms.

At the core of this question is the concept of deterrence and how it relates to cyberattacks and cyberweapons. With regard to *deterrence* as a concept, it is important to note that resulting from the situation and policies of the Cold War, "… the term has acquired not only a special emphasis but also a distinctive connotation" (Brodie, 1959, p. 174), namely that of nuclear deterrence and "… the doctrine of massive retaliation" (Kaufmann, 1954, p. 1). However, even though "[d]eterrence may have

[1] Those and similar terms can be found in the literature spelled with or without a space and sometimes also with a dash instead of a space (e.g. 'cyberattack', 'cyber attack', or 'cyber-attack'; 'cyberdeterrence' or 'cyber deterrence'; 'cyberwarfare' or 'cyber warfare'). In an attempt to keep some consistency, this paper will – outside of direct quotes – use the spelling with neither space nor dash.

assumed a paramount place in the nuclear standoff that the Cold War eventually became …" (Long, 2008, p. 5) it is by itself not a concept that was invented for or because of nuclear weapons (Long, 2008). In fact, the concept of deterrence dates at least as far back as to "… Thucydides … *History of the Peloponnesian War* …" (Long, 2008, p. 5), which shows the long history of the concept of deterrence. The second chapter will provide an overview over this concept, the peculiarities of deterrence using cyberweapons, and how the terms related to it will be used in this paper.

Following that, the game that will form the basis of further analysis will be created based on the scenario implied by the research question as well as deterrence theory and deterrence games as discussed in for example Zagare (2004) and Quackenbush (2011). Afterwards, the fourth chapter will address specific details required to define the last details of the game. For this it will look at three different areas. First, the capabilities of specific types of cyberweapons based on both theoretical examinations and five selected cases that illustrate the destructive and disruptive potential of cyberweapons. Second, the credibility – a core concept of deterrence (Kaufmann, 1954) – of cyberweapons and the escalation potential resulting from credibility (Adamsky, 2013; Gaycken and Martellini, 2013). Third, whether the hypothetical powerful, networked state can be deterred from intervening based on the past performance of cyberweapons in international relations as researched by for example Valeriano and Maness (2015) and research on military interventions conducted by for example Sullivan and Koch (2009). The chapter will conclude with the final definition and a brief analysis of the game. The analysis will be supported by the game theory software Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014).

The game will be simulated by using learning algorithms, specifically "… reinforcement learning (RL) … [and] … belief learning (BL) …" (Moffatt, 2016, p. 420), which are reviewed in the fifth chapter. The sixth chapter will discuss the results of the respective simulations, which were programmed in the programming language Python 3 (Python Software Foundation, no date). The Python 3 code of the simulations is provided in Appendix A as well as online in the GitLab repository belonging to this paper (Gerritzen, 2019), which also contains the data output of the simulations. Selected data tables and additional results can also be found in Appendix B.

The last chapter will be a brief conclusion based the results of the simulations to answer the research question of this paper: Can small states use cyberweapons to deter interventions by powerful, networked states?

2 Deterrence and Cyberdeterrence

2.1 The Concept of Deterrence – Rationality and Fear

2.1.1 The Deterrence Forecast

Morgan (2010, p. 55) defines deterrence as "… efforts to avoid being deliberately attacked by using threats to inflict unacceptable harm on the attacker in response …" to the attack. The literature on deterrence generally distinguishes "… two fundamental approaches to deterrence" (Mazarr, 2018, p. 2). Both approaches will be described in this section before moving on to the concept of deterrence in general to give an overview over what deterrence is.

The first approach is "[*d*]*eterrence by denial*" (Mazarr, 2018, p. 2) which can be summarized as the ability to increase the cost of the attack to the point that it is "… too costly to continue … [or can only result in a] pyrrhic victory" (Morgan, 2010, p. 55). In any case, it directly denies the attacker the potential benefit of the attack (Mazarr, 2018). The second approach is "[*d*]*eterrence by punishment*" (Mazarr, 2018, p. 2). According to Huth and Russett (1988, cited in Mazarr, 2018) strategies that rely on this ability to strike back against the attacker have a lower change of being successful than those of the first type. However, 'deterrence by punishment' "… dominated much of the development of cold-war thinking on deterrence" (Long, 2008, p. 10) and is also less costly than building defense capabilities (Morgan, 2010). Furthermore, the existence of weapons that can circumvent the defenses of the attacker increases the attractiveness of 'deterrence by punishment' strategies, especially when deterrence by denial is not possible (Morgan, 2010).

According to Long (2008, p. 7), "[a] widely used definition of *deterrence* is the manipulation of an adversary's estimation of the cost/benefit calculation of taking a given action" which can be done by either raising the costs or decreasing the benefit of the action in question (Long, 2008). Kaufmann (1954) similarly describes the goal of deterrence as changing the adversary's course of action. In his words, "… deterrence means preventing certain types of contingencies from arising" (Kaufmann, 1954, p. 6) by making "… a forecast about the costs and risks that will be run under certain conditions …" (Kaufmann, 1954, p. 6). This forecast has the goal of changing the adversary's behavior to one that is more favorable to the agent attempting the deterrence strategy (Kaufmann, 1954).

Notably, Kaufmann (1954) states that the forecast, or in other words the threat, does not have to be true; "… it can be in the nature of a bluff" (Kaufmann, 1954, p. 6) and still be effective as long as the adversary believes it to be true (Kaufmann, 1954). This is because, as argued by Morgan (2010, p. 61), "[d]eterrence takes effect in the mind of the opponent – he ultimately determines whether he is deterred". Due to this dependency on the adversary, he argues that "[d]eterrence is a psychological relationship …" (Morgan, 2010, p. 56) which is created with an adversary who is considering actions that the deterrence threat is meant to prevent (Morgan, 2010). This is done by "… shap[ing] an opponent's perceptions, expectations, and ultimately its decisions …" (Morgan, 2010, p. 54). The importance of perceptions is also highlighted by Mazarr (2018). He argues that "[d]eterrence succeeds … by creating a *subjective* perception in the minds of the leaders of the target state" (Mazarr, 2018, p. 7) so that the "… adversary sees the alternatives to aggression as more attractive than war" (Mazarr, 2018, p. 2).

As is already apparent, there is more to deterrence than just threatening the adversary (George and Smoke, 1989); the threat depends on certain "… requirements that a policy of deterrence must fulfill …" (Kaufmann, 1954, p. 8). To start off, for deterrence to work it is "… necessary to surround the proposal with an air of credibility" (Kaufmann, 1954, p. 7). According to Kaufmann (1954, p. 7):

> … [T]here are three main areas in which credibility must be established: the areas of capability, cost, and intentions. The enemy must be persuaded – that we have the capability to act; that, in acting, we could inflict costs greater than the advantages to be won from attaining the objective; and that we really would act as specified …

This again summarizes the basic idea that deterrence is based on a cost/benefit calculation of the adversary which is changed by a threat made against them and highlights that the agent making the threat must have "… the credible capability to harm and the credible intent to carry out this harm" (Long, 2008, p. 8).

Both Kaufmann (1954) and Long (2008) put special emphasis on credibility for effective deterrence, Long (2008, p. 11) even calls it "… the linchpin of deterrence …". However, credibility alone does not lead to an effective deterrence strategy according to George and Smoke (1989) who argue that credibility "… cannot be considered a sufficient condition for deterrence success" (George and Smoke, 1989, p. 177) because historical cases, such as Pearl Harbor, show that, even in case of credible deterrence, attacks may still take place. They further argue that a powerful threat alone is no guarantee for effective deterrence either because "… while 'massive retaliation' was an enormously potent threat, it often lacked enough credibility and relevance …" (George and Smoke, 1989, p. 177) and conclude that both "… credibility and potency of deterrence threat …" (George and Smoke, 1989, p. 177) must be fulfilled for an effective deterrence strategy. In addition to those requirements, Mazarr (2018) identifies two additional conditions.

First, the communication of the threat by the defender (Mazarr, 2018). The importance of communicating the threat is also acknowledged by Kaufmann (1954, p. 6) who stated that "… it becomes necessary to communicate in some way to a prospective antagonist what is likely to happen to him …" if he attacks. Deterrence can not be effective without "… effort to communicate an unambiguous message" (Mazarr, 2018, p. 9) to the adversary.

Second, the "… intentions of the potential aggressor …" (Mazarr, 2018, p. 8). The less important a benefit from a certain action is for an agent, the easier it is to deter that agent from taking that action (Mazarr, 2018). On the other hand, "… if it has acquired an urgent sense that only an attack will safeguard its interest, it may become almost impossible to stop" (Mazarr, 2018, p. 8). In other words, deterrence can only work "… so long as other less intolerable alternative are open …" (Kaufmann, 1954, p. 6) to the potential aggressor; once attacking is seen as the only viable course of action, attempts at deterring attacks will fail (Mazarr, 2018). Related to that is the very similar statement of Brodie (1959, p. 177) who argues:

> … that deterrence has always suggested something relative, not absolute, and that its effectiveness must be measured not only by the amount of power that it holds in check, but also by the incentives to aggression residing behind that power.

In other words, the necessary power of the deterrent depends on the circumstances more so than on the absolute power of either agent (Brodie, 1959). This makes understanding the goals and alternatives of the state that is to be deterred vital to identifying whether deterrence can be effective or not (Mazarr, 2018).

In summary, deterrence is generally defined in the literature as a rational cost/benefit analysis of the adversary that is changed by the credible threat of the deterring agent (Long, 2008). However, despite being "… rooted in thought and calculation, it inherently contains an element of emotion as well" (Long, 2008, p. 7). This element of emotion is "[f]ear [which] exists in the mind of individuals …" (Long, 2008, p. 7) – exactly where deterrence takes place according to Mazarr (2018).

Consequently, at its core, "… [d]eterrence is the generation of fear" (Long, 2008, p. 7). But, as will be shown later, fear is no guarantee for deterrence as it may cause the very thing it is supposed to prevent (Adamsky, 2013)

2.1.2 Deterrence and Superiority

Now that the general ideas behind the concept of deterrence have been addressed it is important to take a brief look at the relationship between deterrence and superiority; is it necessary for one agent to be superior over the other in order to create deterrence? Often, it is assumed that being able to deter an adversary means possessing a superior force with the "… capacity to win a war" (Brodie, 1959, p. 176). Both Brodie (1959) and Mazarr (2018) argue this is not the case. According to Brodie (1959) this is independent of whether nuclear weapons are involved or not even though the presence of nuclear weapons may significantly increase "… the potential deterrence value of an admittedly inferior force …" (Brodie, 1959, p. 177).

Mazarr (2018) additionally makes the point that not just can the inferior agent successfully deter a stronger adversary but that "[s]ometimes states with dominant power refused to fully deploy it …" (Mazarr, 2018, p. 5). In other words, just because an agent is superior it may, for whatever reason, decide to not use its full power against a less powerful agent under all circumstances, which history provides different examples for (Mazarr, 2018). In addition to that, the concept of deterrence itself places "… rais[ing] the cost of a potential attack …" (Mazarr, 2018, p. 6) in the center (e.g. Kaufmann, 1954; Long, 2008; Morgan, 2010). This implies that deterrence is not about winning as "[e]ven if an attacker believes it might be successful … the cost of a long and painful war are a powerful preventive deterrent" (Mazarr, 2018, p. 6).

Brodie (1959) illustrates this with a thought experiment in which the Soviet Union is faced with the possibility "… that a menaced small nation could … [retaliate] with only a single thermonuclear weapon, which, however, it could certainly to deliver on Moscow if attacked" (Brodie, 1959, p. 177) and concludes that this possibility "… would be sufficient to give the Soviet government much pause" (Brodie, 1959, p. 177) despite their certain ability that they would eventually win the war (Brodie, 1959). Therefore, in summary, being able to deter does not imply military superiority in general or even the capability to win against the adversary. This means that, in theory, a small state may very well be able to deter a powerful state given the right conditions and scenario.

2.2 Cyberdeterrence and Cybered Deterrence

2.2.1 Cyberdeterrence

Even though the term *cyberdeterrence* is widely used in the literature, it is lacking a generally accepted definition. Instead, it is used for different ideas that share the similarity of dealing with "… deterrence in cyberspace …" (Rivera, 2012, p. 1) or "[c]yber warfare [which is] … increasingly being recognized as the fifth domain of warfare" (Hughes and Colarik, 2016, p. 20) – the other four domains being "… the physical domains (land, sea, air, and space)" (Philbin, 2013, p. 1). The different concepts will be briefly addressed in this section as well as the following two sections. In addition to there being different ideas of what cyberdeterrence is and what it entails, it is notable that most of the literature is focused on the problems faced by the US and how they can either deter cyberattacks or how they themselves can use them as deterrence (Lupovici, 2011). Neither will be addressed in detail here because it would go beyond the scope of this paper, which takes a different perspective on cyberattacks and deterrence. However, it is still necessary to clarify what is most often meant by cyberdeterrence before looking at the other, very different concept for which the same term may be used (Gaycken and Martellini, 2013; Bendiek and Metzger, 2015).

The concept of cyberdeterrence as it is most often used in the literature deals with the problem of deterring cyberattacks (Keromytis, 2017a). This is also at the core of the definition given by Bendiek and Metzger (2015) who argue that "… cyberdeterrence is built on both deterrence of cyberattacks and deterrence by threatening cyberattacks …" (Bendiek and Metzger, 2015, p. 554) because those aspects "… are different escalatory steps and conceptually cannot be separated" (Bendiek and Metzger, 2015, p. 554). However, this inherent connection between deterring cyberattacks and using cyberattacks as deterrence is debatable. For example, while Rivera (2012) discusses the possibility of deterring cyberattacks with cyberattacks but explicitly states that a "… policy statement aimed at deterring a cyber-attack should include language that does not restrict retaliation to cyberspace alone" (Rivera, 2012, p. 55). This separates the issue of deterring cyberattacks from using cyberattacks as deterrence as both "… a conventional counterattack, or a counterattack through cyberspace …" (Rivera, 2012, p. 48) may be used as deterrence (Rivera, 2012). Keromytis (2017a) states the same and gives a wide range of possible ways to retaliate, "…includ[ing a] cyber 'counterattack' …" (Keromytis, 2017a, p. 53).

2.2.2 Cross-Domain Deterrence

The idea of retaliating against cyberattacks by conventional means described in the previous section is formalized in the rather new and still developing theoretical framework of "cross-domain deterrence (CDD)" (Lindsay and Gartzke, 2016, p. 3) which aims at creating "… a more general theory of means-based deterrence" (Lindsay and Gartzke, 2016, p. 23). According to Lindsay and Gartzke (2016) it differs from "[c]lassical deterrence theory … [where] threats were assumed to be nuclear" (Lindsay and Gartzke, 2016, p. 4) and different means were not addressed by instead "… pay[ing] particular attention to the *means* of deterrence" (Lindsay and Gartzke, 2016, p. 4). Essentially, the theory describes "… the use of threats in one domain … to prevent actions in another domain that would change the status quo" (Lindsay and Gartzke, 2016, p. 6). It is important to note here that the concept of domain used by CDD is different from the use of domain in the previous section. In the context of CDD, it is "… any pathway or means for coercion that is different from other means in important respects …" (Lindsay and Gartzke, 2016, p. 6). For instance, two examples for domains in the context of CDD would "… nuclear and conventional weapons …" (Lindsay and Gartzke, 2016, p. 6).

Even though the concept resulted from concerns about the potential negative effects recent "… developments in space, cyberspace, and other arenas …" (Lindsay and Gartzke, 2016, p. 34) may have on the strategic and geopolitical position of the US and "… whether other American advantages might be brought to bear to compensate" (Lindsay and Gartzke, 2016, p. 34), "… the strategic problem appears more general" (Lindsay and Gartzke, 2016, p. 4). Because of that, the concept itself is not limited to the US as it is basically the policy equivalent to "… combined arms warfare …" (Lindsay and Gartzke, 2016, p. 34) in the sense that strength in one domain compensates for lack of strength in one or multiple other domains (Lindsay and Gartzke, 2016). Additionally, the core question behind the development of CDD is if it "… is fundamentally destabilizing" (Lindsay and Gartzke, 2016, p. 35).

Cyberwarfare by default falls into the area of CDD (Lindsay and Gartzke, 2016). For example, Lewis (2010) argues that deterring cyberattacks requires cross-domain deterrence as retaliation in the same domain would be ineffective. When looking at cyberwarfare and deterrence it is important to make clear that it "… is not a game of great powers …" (Bendiek and Metzger, 2015, p. 566) even though the focus of the majority of the literature is, as also observed by Lupovici (2011), on the US. The next section will elaborate on one way in which it may be used by small states.

2.2.3 Cybered Deterrence

Considering the aforementioned concerns about the strategic and geopolitical position of the US which fostered the development of the concept of CDD, the according to Lindsay and Gartzke (2016) widely found assumption that small states may benefit significantly from recent and upcoming technological developments, and that often "… cyber operations are a result of states trying to match a more powerful opponent … [to] equal the playing field …" (Valeriano and Maness, 2015, p. 55) it makes sense to take a closer look at this concern which manifests in a completely different definition of cyberdeterrence. As stated in the introduction, Rustici (2011), Gaycken and Martellini (2013), and Hughes and Colarik (2016) suggest that small states may be able to use cyberattacks as deterrence. Furthermore, Morgan (2010) suggests that "… cyberattacks will become a great leveler …" (Morgan, 2010, p. 72) if they become easily available because of the "… rising dependence of the U.S. and its friends on cyberspace in many areas …" (Morgan, 2010, p. 72). The potential consequences of this are explored by Rustici (2011) and his idea of cyberdeterrence.

Rustici (2011) outlines that small states could use cyberweapons to deter interventions by powerful, networked states, such as for example the US. He, like Morgan (2010), argues that "… cyberweapons have the potential to become an equalizing force …" (Rustici, 2011, p. 33) and could "… change international relations in ways never seen before" (Rustici, 2011, p. 40). This is, he argues, because they are inexpensive, can strike anywhere, and have significant destructive potential as "[i]n a society as networked as the United States or Europe, most, if not all, of the critical civilian infrastructure is vulnerable to cyberattacks" (Rustici, 2011, p. 34). Consequently, "[n]etworked societies will be far more cautious in advocating for humanitarian intervention, regime change, no-fly zones, and other nonessential security operations" (Rustici, 2011, p. 40). This places the focus on "… deterrence by threatening cyberattacks …" (Bendiek and Metzger, 2015, p. 554) which Rustici (2011) states "… is truly defense on the cheap" (Rustici, 2011, p. 40). Notably, his definition of cyberdeterrence does not include deterring cyberattacks at all, which illustrates the definition issue of the term cyberdeterrence quite well.

Due to this definition problem it makes sense to follow Gaycken and Martellini (2013, also cited in Bendiek and Metzger) who distinguish two different terms, which both have a specific meaning. They use the term “cyber deterrence” (Gaycken and Martellini, 2013, p. 1) when referring to the deterrence *of* cyberattacks. When it comes to the use of cyberattacks *for* deterrence they do not use the term cyberdeterrence, unlike Rustici (2011). Instead, they use the term “cybered deterrence” (Gaycken and Martellini, 2013, p. 1). This is also how the terms will be used in this paper in order to clearly distinguish the two different concepts that have been established as they both come with their own, very different goals Gaycken and Martellini (2013).

Despite their differences, cybered deterrence, just like cyberdeterrence, can also be considered an example for CDD because “[i]n-kind deterrence will not be an aim of cybered deterrence” (Gaycken and Martellini, 2013, p. 6). This leads back to the statement of Lindsay and Gartzke (2016) that cyberwar and CDD are by default connected. Cybered deterrence is an example for CDD because an actor is “… driven by overwhelming conventional inferiority” (Rustici, 2011, p. 40) to achieve deterrence through the domain of cyberspace by using “[o]ffensive cyber capabilities … [which are] a new and … unconventional kind of deterrent” (Gaycken and Martellini, 2013, p. 1) that is an additional method of deterrence (Gaycken and Martellini, 2013). This perfectly fits the definition of CDD provided by Lindsay and Gartzke (2016, p. 6) where “… threats in one domain … [are used] to prevent actions in another domain that would change the status quo”, the status quo here being the absence of an intervention.

2.2.4 Doctrines of Cybered Deterrence

Just like nuclear deterrence is characterized by “… the doctrine of massive retaliation” (Kaufmann, 1954, p. 1) and “… mutually assured destruction …” (Iasiello, 2014, p. 55) there are a number “… of possible doctrines of cybered deterrence …” (Gaycken and Martellini, 2013, p. 4) that can be used by states that want to employ cybered deterrence (Gaycken and Martellini, 2013). While cybered deterrence has yet to become reality, it is nonetheless possible to develop potential doctrines that may be used in the future (Gaycken and Martellini, 2013). Due to the large number of possible operations in cyberspace there are also a number of possible deterrence doctrines, which operate in different ways and require different capabilities. Six possible doctrines are described in detail by Gaycken and Martellini (2013). Those doctrines will be briefly summarized in the following and will be used to support the analysis of cybered deterrence against humanitarian military interventions.

The first two doctrines described by Gaycken and Martellini (2013) are related to the scope of the capability. Thus, the doctrine can be one "… of 'Targeted Capability'" (Gaycken and Martellini, 2013, p. 4) or "General Capability" (Gaycken and Martellini, 2013, p. 4). The difference is that in the case of the 'Targeted Capability' doctrine, the agent "… would demonstrate only a specific capability to attack very specific systems [such as] … financial software used at stock exchanges" (Gaycken and Martellini, 2013, p. 4) whereas a 'General Capability' doctrine requires the demonstrated "… ability to hack all kinds of systems …" (Gaycken and Martellini, 2013, p. 4).

Independent of whether a 'Targeted Capability' or 'General Capability' doctrine is chosen, there are different methods that can be used to implement either (Gaycken and Martellini, 2013). Those methods – or rather types of cyberattacks – have been categorized into three groups by Taipale (2009 cited in Rivera, 2012) – "… cyber-espionage, … data disruption, … and cyber isolation" (Rivera, 2012, p. 10). Those three groups translate well into the two potential main doctrines Gaycken and Martellini (2013) have developed for the implementation of 'Targeted Capability' and 'General Capability' which are the "Assured Disruption" (Gaycken and Martellini, 2013, p. 5) doctrine and the "Forced Transparency" (Gaycken and Martellini, 2013, p. 5) doctrine.

A 'Forced Transparency' doctrine would be aimed at threatening the secrecy of the adversary's secrets (Gaycken and Martellini, 2013). In cyberspace, this would be an example for cyberespionage (Rivera, 2012). In more general terms, one may also call it a "Computer Network Exploitation (CNE)" (Rivera, 2012, p. 4; Philbin, 2013, p. 12) which "… readily equates to spying in the physical domains" (Philbin, 2013, p. 12). It may also be used to identify "… vulnerabilities that may be exploited later" (Philbin, 2013, p. 12) to execute a "Computer Network Attack (CNA)" (Rivera, 2012, p. 4; Philbin, 2013, p. 11).

An 'Assured Disruption' doctrine "… would demonstrate an ability to disrupt vital IT-services or data streams …" (Gaycken and Martellini, 2013, p. 5). This covers the attacks of the types 'data disruption' and 'cyber isolation' which respectively aim at the "… disruption or manipulation of information systems or infrastructure" (Rivera, 2012, p. 11), including "… critical infrastructure … [such as] electricity … or water supplies" (Rivera, 2012, p. 12), or the provision of online services (Rivera, 2012). The latter includes "Distributed Denial of Service (DDoS)" (Rivera, 2012, p. 12) attacks which are, as argued by Philbin (2013), also a type of CNA. Furthermore, the "… simple disruption of services …" (Gaycken and Martellini, 2013, p. 5) is included in 'Assured Disruption' doctrine (Gaycken and Martellini, 2013). This is because a DDoS could be a "… strategic threat" (Rivera, 2012, p. 12) if it "… occurs over an extended time period and prevents access to critical parts of either service or economic infrastructures …" (Rivera, 2012, p. 12); to say it in the words of Gaycken and Martellini (2013, p. 5): "It is the wooden club in the armory of the cyber solider. But a very large wooden club will have a deterrent effect nonetheless".

The doctrines described so far focused on the execution of powerful, individual attacks against the adversary (Gaycken and Martellini, 2013). However, this is not the only way to cybered deterrence as the doctrine of "Silent Erosion" (Gaycken and Martellini, 2013, p. 5) illustrates. The 'Silent Erosion' doctrine may be "… the easiest and most worrisome ability among all cyber abilities [because it] … weaken[s] and slowly erod[es] the society targeted" (Gaycken and Martellini, 2013, p. 5) through numerous small attacks (Gaycken and Martellini, 2013) that for the adversary may be "… impossible to defend against" (Gaycken and Martellini, 2013, p. 5). The general idea is similar to the suggestion that there may be "… psychological cyber-attack effects…" (Rivera, 2012, p. 11). This suggestion can also be found in Rustici (2011) who attributes to cyberweapons a "… distinctive psychological impact … [that] cannot be underestimated" (Rustici, 2011, p. 40), which results from the impossibility to create adequate defenses as well as the uncertainty about the consequences and timing of a potential attack (Rustici, 2011). The same argument is made by Gaycken and Martellini (2013) in the context of the 'Silent Erosion' doctrine.

The last potential doctrine that is introduced by Gaycken and Martellini (2013) is the "Digital Media Control" (Gaycken and Martellini, 2013, p. 5) doctrine and its special form, the "Attribution Control" (Gaycken and Martellini, 2013, p. 5) doctrine. Both doctrines use the ability to control information in a way that harms the adversary. In particular the 'Digital Media Control' doctrine does not depend on the ability to conduct actual cyberattacks (Gaycken and Martellini, 2013). Instead, it is based on the ability to "… spin information operations and the knowledge of how to place these effectively in digital media" (Gaycken and Martellini, 2013, p. 5). 'Attribution Control' as a special form of 'Digital Media Control' goes a step beyond spinning information to harm the adversary (Gaycken and Martellini, 2013). This doctrine weaponizes the – in literature on cyberdeterrence often discussed – "attribution problem" (Lindsay, 2015,

p. 53), which may cause "… uncertainty about the very identity of the opponent" (Lindsay, 2015, p. 56), by using "… false-flag operations" (Gaycken and Martellini, 2013, p. 5). Those false-flag operations would deter the adversary by causing "… fear that the deterring party would always be capable to escalate tensions between the defender and a third party" (Gaycken and Martellini, 2013, pp. 5–6).

In conclusion, while cybered deterrence may lack "… the same deterrent value as nuclear deterrence" (Gaycken and Martellini, 2013, p. 9) it may still be valuable for small and large states alike (Gaycken and Martellini, 2013).

3 Deterrence Theory and Deterrence Games

3.1 Classical and Perfect Deterrence Theory

Keeping in mind the concepts described in the previous parts of this paper, this chapter will develop the basic structure of the game that will be analyzed. It will also determine the aspects that need to be considered when defining the parameters of the game and interpreting the results of the simulations.

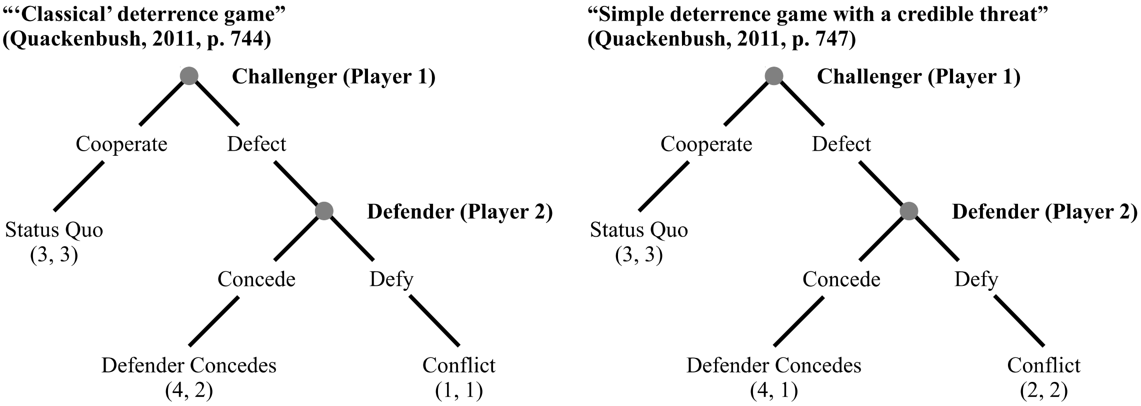
The first step to use simulations to answer the research question is to create a game that represents a simplified model of the international system; simulating reality would be far too complex. This model will take the form of a sequential game with two players, each representing one state, which puts the game theoretic modeling of a deterrence scenario at the center of this paper. Therefore, it makes sense to take a closer look at deterrence games which are game theoretic representations of deterrence scenarios (Quackenbush, 2011). Deterrence games are commonly used in the literature when modeling deterrence scenarios or discussing their theoretical foundation, deterrence theory, as can be seen in for example Zagare (2004) or Quackenbush (2011). This section will take a look at both deterrence theory and deterrence games in order to create the foundation for the game that will be developed and analyzed throughout this thesis.

Whereas Morgan (2003, cited in Quackenbush, 2011, p. 743) argues that "… there may be different deterrence *strategies* [but] there is only one deterrence *theory*", Quackenbush (2011) identifies two different strands of deterrence theory, which "… both are rational-choice theories …" (Quackenbush, 2011, p. 761), in his examination of deterrence theory. Those theories are "… classical deterrence theory and perfect deterrence theory" (Quackenbush, 2011, p. 761). Within classical deterrence theory two sub-theories can be identified according to Zagare (2004 and 1996, cited in Quackenbush, 2011, p. 743) which are "… *structural deterrence theory* and *decision-theoretic deterrence theory*".

To give a short summary, structural deterrence theory builds on the assumptions of the International Relations theory of Realism (Zagare, 2004; Quackenbush, 2011). They argue "… that the key to international stability lies in the distribution of power in the international system and the absolute cost of war" (Zagare, 2004, p. 109) and that therefore "… war becomes unthinkable (i.e. irrational) once power is balanced and the cost of war is exorbitant" (Zagare, 2004, p. 110) because of the "… monotonic relationship between the cost and probability of war" (Zagare, 2004, p. 110). Based on that they reach "… *the* central conclusion of structural deterrence theory: that war in the nuclear age is 'irrational'" (Zagare, 2004, p. 111).

Decision-theoretic deterrence theory "… can be seen as a micro- (or unit-)level extension of structural deterrence theory …" (Zagare, 2004, p. 112) because it adopts the aforementioned conclusion of structural deterrence theory as basic assumption; the theory "assum[es] that conflict is the worst outcome for both players … [, therefore] presum[ing] war to be irrational" (Zagare, 2004, p. 112). However, there is a problem with this theory when it comes to its implications. The assumption that war is always the worst outcome in combination with rational-choice by both players leads to the conclusion that deterrence can never be credible as no player would ever follow through with any threat (Zagare, 2004; Quackenbush, 2011).

Figure 1: Classical and Simple Deterrence Game (Quackenbush, 2011)



"'Classical' deterrence game" (Quackenbush, 2011, p. 744)

"Simple deterrence game with a credible threat" (Quackenbush, 2011, p. 747)

Own illustration, based on "'Classical deterrence' game" (Quackenbush, 2011, p. 744) and "Simple deterrence game with a credible threat" (Quackenbush, 2011, p. 747).

This is illustrated by the "'[c]lassical' deterrence game" (Quackenbush, 2011, p. 744) which can be found in Figure 1 on the left. In this game it can be demonstrated by “… using *backwards induction* [which means that] … one works backwards up the game tree …” (Zagare, 2004, p. 113) to identify the decisions rational agents will make. First, from the perspective of the defender, the strategy ‘Concede’ dominates ‘Defy’ which means “… an instrumentally rational … [defender] will choose to concede …” (Zagare, 2004, p. 114). Second, we assume that deterrence is necessary which in other words means “… [the challenger] has an incentive to upset the *Status Quo* and … will rationally choose [‘Defect’]” (Zagare, 2004, p. 114). From that it follows that “… the *Status Quo* is unstable and deterrence rationally fails” (Zagare, 2004, p. 114) as it would require irrational behavior of the defender to be credible (Zagare, 2004; Quackenbush, 2011). Because of this problem, (Quackenbush, 2011) states that “… scholars need to move away from the assumption that conflict is the worst possible outcome” (Quackenbush, 2011, p. 762) as done by Zagare and Kilgour (2000, cited in Quackenbush, 2011, p. 762) who developed “… perfect deterrence theory [(PDT), which] provides a logically consistent alternative to understand the dynamics of deterrence”. The name of the theory is derived from its creators’ “… insistence on the use of perfect equilibria” (Quackenbush, 2011, p. 747). *Perfect* refers to “… equilibri[a] … [which are] sub-game perfect” (Quackenbush, 2011, p. 747).

Even though there are similarities, – in both theories “… states are assumed to be rational and egoistical” (Zagare, 2004, p. 117) – perfect deterrence theory differs from classical deterrence theory (Zagare, 2004). Perhaps most importantly, it removes, as stated before, the assumption of classical deterrence theory “… that conflict is always the worst possible outcome …” (Quackenbush, 2011, p. 762) and argues that “… only rational (i.e. credible) threats can be carried out” (Zagare, 2004, p. 118). This is illustrated by the “[s]imple deterrence game with a credible threat” (Quackenbush, 2011, p. 747) shown in Figure 1 on the right, which is based on perfect deterrence theory (Quackenbush, 2011). In this game, the defender is making a credible threat, which can be executed (Zagare, 2004; Quackenbush, 2011). Here, it is important to note that perfect deterrence theory distinguishes capability and credibility (Zagare, 2004). As explained, a threat is credible if the strategy to execute the threat dominates the alternative (Zagare, 2004). In the context of the games in Figure 1, this means that player 2, the defender, prefers ‘Defy’ over ‘Concede’ (Quackenbush, 2011). However, “… in the absence of a necessary condition (i.e. a capable threat), a credible threat is insufficient for ensuring deterrence success” (Zagare, 2004, p. 125). This introduces the requirement of capability; a “… threat is capable only if the other, the *threatened* player, prefers the status quo to the outcome when and if the threat is carried out” (Zagare, 2004, pp. 123–124). At the same time this also implies that, no matter how credible or powerful the threat is, deterrence cannot work if the challenger prefers ‘Conflict’ over ‘Status Quo’ despite the threat (Zagare, 2004).

Furthermore, this approach to deterrence theory is also consistent with the findings of various authors that have been summarized earlier (Section 2.1); it formalizes the different observations in a unified model and also appears to be “… consistent with the empirical records (Quackenbush and Zagare, 2001; Quackenbush, 2003; Senese and Quackenbush, 2003)” (Zagare, 2004, p. 117). There are different relevant aspects of the concept of deterrence from previously cited literature.

First, the finding of George and Smoke (1989) who made a similar argument as perfect deterrence theory with regard to the requirement of capability and credibility and stated that “… some combination of these two conditions – credibility and potency of deterrence threat – is relevant for deterrence” (George and Smoke, 1989, p. 177).

Furthermore, different authors argued that “[d]eterrence takes effect in the mind of the opponent – he ultimately determines whether he is deterred” (Morgan, 2010, p. 61). Similar ideas can be found in for example Kaufmann (1954) or Mazarr (2018). This, together with the argument that “… the intentions of the potential aggressor …” (Mazarr, 2018, p. 8) and “… the degree to which a potential aggressor is dissatisfied with the status quo …” (Mazarr, 2018, p. 8) are of significant importance (Mazarr, 2018), are also considered by perfect deterrence theory as outlined by Zagare (2004). The theory includes the possibility that “… threats may lack capability if the threatened state calculates that the cost of conflict is less than the cost of doing nothing” (Zagare, 2004, p. 124). This also refers back to the “… widely used definition of *deterrence* [according to which it] is the manipulation of an adversary’s estimation of the cost/benefit calculation of taking a given action” (Long, 2008, p. 7).

Lastly, it is important to note that perfect deterrence theory is – as opposed to classical deterrence theory – not focused on nuclear deterrence (Zagare, 2004). Instead, it is “… a universal theory of conflict initiation and resolution, applicable to both nuclear and to non-nuclear interactions” (Zagare, 2004, p. 134). This is of particular importance here given that the question of this paper deals with non-nuclear deterrence.

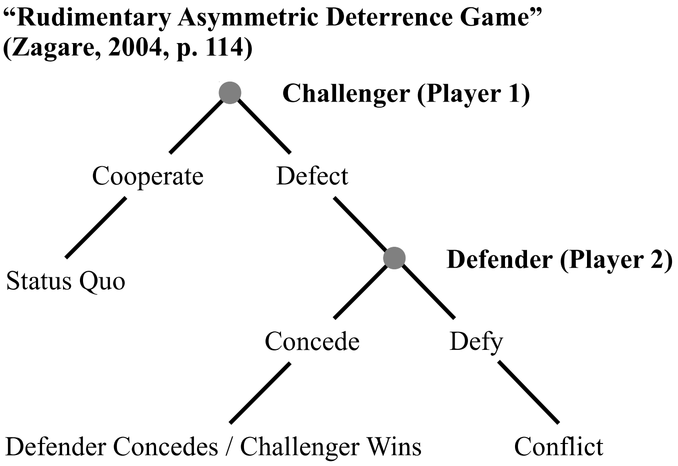
Those aspects taken together point at perfect deterrence theory being a solid theoretical foundation for the ‘Sequential Cybered Deterrence Game’ that will be drawn up in the following parts of this paper. Thus, this paper follows the advice of

Quackenbush (2011) who argued in favor of using perfect deterrence theory, stating that it "… provides the most appropriate basis for further theoretical development, empirical testing, and application to policy" (Quackenbush, 2011, p. 762).

3.2 Rudimentary Asymmetric Deterrence Game

Two different deterrence games have already been shown in the previous section, the "'[c]lassical' deterrence game" (Quackenbush, 2011, p. 744) and the "[s]imple deterrence game with a credible threat" (Quackenbush, 2011, p. 747), which both share the same constellation of actors and strategies. They can be generalized as the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114) which is described by Zagare (2004) as "… a model of an asymmetric or one-sided deterrence situation …" (Zagare, 2004, p. 113). This generalized game is shown in Figure 2.

Figure 2: Rudimentary Asymmetric Deterrence Game (Zagare, 2004)



Own illustration, based on "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114).

According to Zagare (2004, p. 113) this game is "… perhaps the simplest deterrence situation that one can imagine …". As can be seen in Figure 2, it features a challenger and a defender who both can choose between two different strategies (Zagare, 2004). The challenger can either 'Cooperate' or 'Defect' whereas the defender can either 'Concede' or 'Defy' (Zagare, 2004). Because the game is a sequential game, it starts with the decision of the challenger who can either "… accept the status quo or … *defect*" (Zagare, 2004, p. 113). Then, the defender can either 'Concede' and let the challenger win or 'Defy', leading to 'Conflict' between the two players (Zagare, 2004).

As a final point, the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114) shares its basic structure with the – slightly more complex – "Unilateral Deterrence Game" (Zagare, 2004, p. 119). Both types of games feature "… players

[that] have distinct roles and distinct motivations: one player … hopes to preserve the status quo while the other … would prefer to overturn it" (Zagare, 2004, p. 122).

3.3 Sequential Cybered Deterrence Game

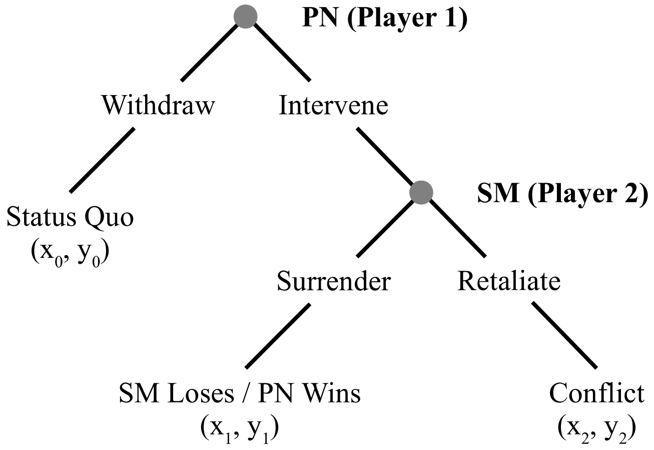
The structure of the 'Sequential Cybered Deterrence Game' (SCDG) will be based on the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114) shown in Figure 2 as well as partly on the "Unilateral Deterrence Game" (Zagare, 2004, p. 122), which have been briefly addressed in the previous sections. The "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114) in particular has been chosen because its constellation of actors and strategies represents the scenario that is the topic of this paper – the hypothetical scenario will be described in more detail in this section – while remaining a relatively simple game to analyze and simulate.

Just like in the "… Unilateral Deterrence Game, the players have distinct roles and distinct motivations" (Zagare, 2004, p. 122) in the SCDG as well. In fact, they have the same dynamic of a "… 'Defender, [who] hopes to preserve the status quo … [and a] 'Challenger', [who] would prefer to overturn it" (Zagare, 2004, p. 122) that can be found in both of the referenced games that have been taken from Zagare (2004). To make the players more specific, they will be given two distinct identifiers here. The 'Challenger' will be the powerful, networked state (state PN) whereas the 'Defender' is a small state (state SM).

Since the analysis is based around a game – a sequential game will be used as a starting point with the strategic normal form being analyzed later in addition as well – the interaction between the two states or players will be limited to the available strategies in the game. While being based on the game described by (Zagare, 2004), the game in question is also based on a hypothetical scenario. This scenario is not directly based on any particular real case and does not involve any real state, even though similarities are inevitable when addressing the question of deterrence against powerful, networked states. The capabilities, vulnerabilities, and past or expected behavior of real states will be used to determine the payoffs of the players in the simulation and to aid the interpretation of the simulation outcomes. The purpose of the scenario is to make to analysis less abstract and to give a possible context for cybered deterrence.

In the scenario, state SM is subject to an unspecified internal crisis. In response to this crisis and continued severe large scale human rights violations by state SM's government, state PN has urged state SM to cease the violence against its citizens and adhere to its responsibilities under international law. Attempts to force state SM's compliance through measures such as economic sanctions have failed to create the intended response. In addition to that, state SM has consistently threatened to retaliate with cyberattacks if military action of any kind is taken against them. In the past, state PN has used its military capabilities to intervene in such internal crises, often forcing a change of government in the process. Given this situation, state PN has now to decide whether to intervene or not. In summary, the scenario represents the attempt of a small state to use cybered deterrence – as a "[*d*]*eterrence by punishment*" (Mazarr, 2018, p. 2) strategy – against a powerful, networked state to prevent a military humanitarian intervention and potential regime change, a scenario like it has been described by (Rustici, 2011).

Figure 3: Sequential Cybered Deterrence Game (undefined payoffs)



Own illustration, adapted from the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114).

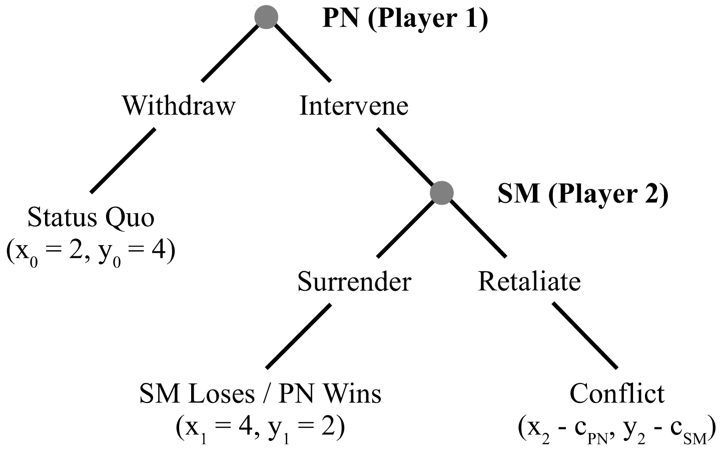
As can be seen in Figure 3, the available strategies and respective outcomes are essentially the same as those of the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114). The state PN makes the first move and decides whether to 'Intervene' (I) or to 'Withdraw' (W) from the (potential) open conflict with state SM. This is the same decisions as that of the 'Challenger' to 'Cooperate' or to 'Defect' that can be found in the game of Zagare (2004, p. 114) or the games that can be found in Quackenbush (2011, p. 744 and 747). If state PN decides to 'Intervene', state SM has to decide whether to 'Surrender' (S) or to 'Retaliate' (R) which mirrors the decision of the 'Defender' between 'Concede' or 'Defy' in the games of Zagare (2004, p. 114) and Quackenbush (2011, p. 744 and 747). The payoffs of the outcomes are labeled x_0, x_1, and x_2 for state PN and y_0, y_1, and y_2 for state SM. The payoffs x_2 and y_2 – the payoffs for the outcome 'Conflict' – will be influenced by different factors related to cybered deterrence and its doctrines as they are described by Gaycken and Martellini (2013) as well as perfect deterrence theory. This will be addressed in more detail in the following section which will take a look at the definition of the payoffs for the SCDG.

3.4 Defining the Payoffs of the Sequential Cybered Deterrence Game

This section will address the definition of payoffs of the SCDG in general, the following chapter will – based on the results of this section – address specific issues and conditions. As a base line for the payoffs this paper will adopt the scale used by Quackenbush (2011) for both the "'[c]lassical deterrence game" (Quackenbush, 2011, p. 744) and the "[s]imple deterrence game with a credible threat" (Quackenbush, 2011, 2p. 747). The scale ranges from 1 for the worst payoffs to 4 for the best payoffs (Quackenbush, 2011). In addition to that, the model will also draw from Fey and Ramsay (2011) who included in their model that "[e]ach country also pays a cost of war" (Fey and Ramsay, 2011, p. 154). Here, the variable cost of conflict will be represented by the variables c_{PN} and c_{SM} for the respective state and a fixed cost will be included in the base payoffs x_2 and y_2.

Based on the described scenario, the payoff scale of Quackenbush (2011), and the inclusion of cost of conflict, as also done by Fey and Ramsay (2011), it is possible to define the payoff of the outcomes 'Status Quo' and 'SM Loses / PN Wins' since neither is impacted by cybered deterrence. Both outcomes lead to the best payoff for one of the players. In case of the outcome 'Status Quo', it is the best outcome for state SM as no intervention takes place or state PN ends their intervention unsuccessfully. Opposed to that is 'SM Loses / PN Wins' which is the best outcome for state PN as their intervention ends successfully and thus they achieve their goal. For both outcomes, the respective other player is given the "next-worst" (Quackenbush, 2011, p. 744) payoff to which Quackenbush (2011) has assigned the value 2. This is done because it is neither a particularly favorable nor the worst possible outcome; both the "next-best" (Quackenbush, 2011, p. 744) and "worst" (Quackenbush, 2011, p. 744) payoff remains a possibility for 'Conflict' as it could be either better or worse than the alternative. For example, a costly conflict with disastrous consequences may be a worse alternative for both than avoiding the conflict. Figure 4 shows the game with those payoffs assigned.

Figure 4: Sequential Cybered Deterrence Game (variable 'Conflict' payoff)



Own illustration, adapted from the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114), the "'Classical' deterrence game" (Quackenbush, 2011, p. 744), the "Simple deterrence game with a credible threat" (Quackenbush, 2011, p. 747), and the "Simple Crisis Game" (Fey and Ramsay, 2011, p. 154).

In addition to the already assigned payoffs, it is further possible to define a set of four conditions for the payoff of the outcome 'Conflict'. Two of those conditions must be fulfilled independent of the success of deterrence whereas two – according to perfect deterrence theory as described by Zagare (2004) – must be fulfilled additionally in order to have effective deterrence. The first two conditions are derived from the cost of conflict, which implies that the payoff of the outcome 'Conflict' may never be as good as the payoff of outcomes that do not have the cost of conflict (Fey and Ramsay, 2011).

1. For state SM this implies that y_2 - c_{SM} < y_0, i.e. the outcome 'Conflict' may never result in a higher payoff than 'Status Quo'.
2. For state PN this implies that x_2 - c_{PN} < x_1, i.e. the outcome 'Conflict' may never result in a higher payoff than 'SM Loses / PN Wins'.

While the above conditions are fulfilled by design, the second set of conditions are based on capability and credibility in the sense of perfect deterrence theory. As the terms are used for specific concepts in the context of perfect deterrence theory, they will be written as 'capability (PDT)' and 'credibility (PDT)' to distinguish them from the general usage of both words as they appear quite often when talking about deterrence.

3. According to perfect deterrence theory, a "… threat is capable only if the other, the *threatened player*, prefers the status quo to the outcome that results when and if the threat is carried out" (Zagare, 2004, pp. 123–124). This implies that the capability (PDT) requirement for deterrence is fulfilled if x_2 - c_{PN} < x_0, i.e. if the payoff of 'Conflict' is lower than the payoff of 'Status Quo'.

4. Lastly, there is the credibility (PDT) requirement (Zagare, 2004). It is only fulfilled for threats are "… that the threatener prefers to execute" (Zagare, 2004, p. 125). This means that that threats are credible when y_2 - $c_{SM} > y_1$, i.e. when the payoff of 'Conflict' is higher than the payoff of 'SM Loses / PN Wins'.

In order to identify whether there is capability (PDT) and credibility (PDT), it makes sense to look in particular at three different aspects. First, the potential damage that could be inflicted on state PN by the deterrent (Zagare, 2004). Second, whether state PN is convinced that the deterrent is indeed real (Kaufmann, 1954). Third, the impact of this threat and/or damage on the behavior of state PN which heavily influences capability (PDT) (Zagare, 2004; Mazarr, 2018).

4 Capability (PDT) and Credibility (PDT) of Cybered Deterrence

This chapter will address in detail whether there is capability (PDT) and credibility (PDT) and will briefly analyze the fully defined game. It is structured into four main sections, one for each of the three aspects that influence capability (PDT) and credibility (PDT) as well as one for the final definition of the simulation parameters and the analysis of the game, which will mark the end of this chapter.

First, to determine the potential damage, selected cases of cyberattacks, including cyberattacks conducted by non-state actors, as well as analyses about the vulnerability of the US – a good example of an PN-type state, considering the focus of, among others, Rustici (2011, p. 40) on "… the United States and other advanced states …" –, general research into the vulnerability of systems, and potential cyberweapons will be considered. Such considerations can also be commonly found in literature on cyberdeterrence, such as for example Rivera (2012), which is, as mentioned before, mostly "… based on an American perspective" (Lupovici, 2011, p. 49).

The second section will then take a look at the "… credibility issue …" (Valeriano and Maness, 2015, p. 59) of cyberdeterrence, which is in particular related to the "… credibility of capability …" (Long, 2008, p. 11), and how cybered deterrence may share the same issue. This is also addressed by Gaycken and Martellini (2013) in the context of building its "… force posture …" (Gaycken and Martellini, 2013, p. 4). However, credibility does not come without cost. Adamsky (2013, p. 33) argues that there is a "… 'culminating point of deterrence' … [beyond which] credible threats become so convincing that …" they may cause a first strike.

Similarly, Gaycken and Martellini (2013, p. 4) state that "… the deterrence posture … has to be crafted very carefully to avoid misperceptions and unintended escalation" and further note the different potential of escalation that comes with the different doctrines of cybered deterrence they describe. Therefore, it is important to briefly consider the escalation potential of cybered deterrence.

Lastly, since the importance of the adversary and its position has been highlighted by different authors (e.g. Brodie, 1959; Mazarr, 2018; see in particular section 2.1 and 3.1), a look will be taken at the perspective of state PN to identify whether state SM can deter the intervention. So far there appears to be no known case of cybered deterrence; Gaycken and Martellini (2013) purely consider it as a potential method of deterrence. Because of that, past cyberattacks in international relations and reactions to resistance against interventions which increased their cost will be used instead.

4.1 Defining Cyberweapons and Identifying their Capabilities

The term *cyberweapon* is often found in the literature and while it may seem straightforward at first glance, there is a lot of debate about the properties of cyberweapons. While it is possible to "… define and profile conventional weapons … or unconventional weapons like … chemical, biological, nuclear or radiological and improvised weapons …" (Maathuis, Pieters and Den Berg, 2016, p. 1) this is not yet the case with cyberweapons (Maathuis, Pieters and Den Berg, 2016). This is because they "… are an uncertain concept due to the fact that there is no accepted global definition and there is a lack of research concerning their profile, action and impact" (Maathuis, Pieters and Den Berg, 2016, p. 1). In other words, there is a lack of agreement on what cyberweapons are and what they are capable of.

This lack of agreement also manifests in the literature which can be illustrated by looking at two perspectives on cyberweapons. On the one end of the spectrum there is the statement of Valeriano and Maness (2017) who see cyberweapons as "… complicated, expensive, and difficult to utilize for offensive and defense intent" (Valeriano and Maness, 2017, p. 261) even though they stated before that "… cyber power is cheaper to acquire than military hardware" (Valeriano and Maness, 2015, p. 25). On the other end of the spectrum we can find for example Rustici (2011) who states that they "… are cheap, effective, and can be utilized from anywhere in the world, at any time" (Rustici, 2011, p. 36) while it is also "… very easy to develop this [cyberattack] capacity with an exceedingly small footprint" (Rustici, 2011, p. 38).

To start off, it makes sense to take a look at the definition of the term cyberweapon that can be found in the English language Oxford Dictionaries according to which a cyberweapon is “[a] piece of computer software or hardware used to commit cyberwarfare” (*cyberweapon | Definition of cyberweapon in English by Oxford Dictionaries*, no date) where cyberwarfare – a term that is the subject of much debate in the literature (Valeriano and Maness (2015) – is “[t]he use of computer technology to disrupt the activities of a state or organization, especially the deliberate attacking of information systems for strategic or military purposes” (*cyberwarfare | Definition of cyberwarfare in English by Oxford Dictionaries*, no date). This simple definition already contains many elements that will also be part of the following three definitions. Notably, any indication of the exact capabilities of the weapons is absent in this definition.

The first definition of cyberweapon from the literature that will be addressed is that of Crowther (2017), which can be found in the *Encyclopedia of Cyber Warfare* (Springer, 2017). He defines it as a “… term used to describe programs, equipment, tactics, techniques and procedures used for offensive cyber operations” (Crowther, 2017, p. 76). He further distinguishes “… two types of cyber attacks: semantic and syntactic … [as well as] two types of effects that a cyber attack can achieve: manipulation and denial” (Crowther, 2017, p. 76). While “[s]emantic attacks use language …[, s]yntactic attacks use computer codes …” (Crowther, 2017, p. 76). Crowther (2017) further states that an attack may use both types. For example, “… the first phase of a phishing attack is a semantic attach where the attacker convinces the target to click on the link [whereas] the second, or syntactic, phase of the attack unleash[es] the malware into the target system” (Crowther, 2017, p. 76). With regard to the aforementioned two types of effects, “[m]anipulation describes any change [of] … the thoughts of the target … [or] of coding” (Crowther, 2017, p. 76). ‘Denial’ further has three sub-types, which are “… degradation, disruption, and destruction” (Crowther, 2017, p. 76). Lastly, Crowther (2017) describes a number of types of specific methods of attack, illustrating how different cyberweapons can be.

Another definition and way to group cyberweapons is used by Valeriano and Maness (2015). Here, they are defined as “computer codes that are used, or designed to be used, with the aim of threatening or causing physical, functional, or mental harm to structures, systems, or living beings” (Rid and McBurney, 2012, p. 6 cited in Valeriano and Maness, 2015, p. 33). This definition is much more narrow than the previous definition as it is limited to ‘computer codes’ and does not include other, related aspects found in the definition of Crowther (2017). Based on this definition, Valeriano and Maness (2015) identify "… four basic methods (weapons) that cyber conflict initiators have at their disposal" (Valeriano and Maness, 2015, pp. 33–34). Those four types are:

> … [1)] website defacement and vandalism … [, 2)] DDoS [attacks] … [, 3)] [i]ntrusions, which includes Trojans and … backdoors … [, 4)] infiltrations. … Infiltrations and intrusions are not scalar with regard to which one is more severe, but they are generally more sophisticated [and] … targeted … (Valeriano and Maness, 2015, pp. 34–35).

As can be seen above, Valeriano and Maness (2015) basically use the method of attack to group cyberweapons instead of using their effect as suggested by Crowther (2017). However, both illustrate – just like the number of cybered deterrence doctrines described by Gaycken and Martellini (2013) – the wide range of cyberweapons.

The last definition that will be addressed here is the definition of Maathuis, Pieters and Den Berg (2016, p. 4) who:

> … propose the following definition for a cyber weapon:
> **A computer program created and/or used to alter or damage (an ICT[[2]] component of) a system in order to achieve (military) objectives against adversaries inside and/or outside cyberspace.**

This definition is similar to the definition of Rid and McBurney (2012, also cited in Valeriano and Maness, 2015) but differs in three important aspects. First, it does not include "… the aim of threatening … harm …" (Rid and McBurney, 2012, p. 6 cited in Valeriano and Maness, 2015, p. 33) which Rid and McBurney (2012) argue is of significant importance. This is because "… a tool is actually used as a weapon when an actor *is intending to use it as such*; whether harm is successfully inflicted or not is of secondary concern" (Rid and McBurney, 2012, p. 7). Second, the definition of Maathuis, Pieters and Den Berg (2016) includes an additional aspect, which is a clear definition of the target, that being "… an ICT system e.g. application, data, device, or … non-ICT system that contains an ICT component that represents practically the carrier to the desired target" (Maathuis, Pieters and Den Berg, 2016, p. 4).

Just like the first two definitions, the definition of Maathuis, Pieters and Den Berg (2016) is a general definition that does not make any assumptions about what exactly a cyberweapon looks like. In fact, they describe in detail a variety of "… characteristics and classification criteria of cyberweapons …" (Maathuis, Pieters and Den Berg, 2016, p. 1) that can be used to "… analyse and profile cyber weapons …" (Maathuis, Pieters and Den Berg, 2016, p. 5). This means that it is not practical to make generalized

[2] ICT is the abbreviation for "[i]nformation and communications technology" (ICT | Definition of ICT in English by Oxford Dictionaries, no date).

assumptions about the usefulness and potential capability of cyberweapons. Instead, it seems to be more appropriate to take a look at a number of specific examples.

In particular, three specific types of cyberattacks will be considered in the following three sub-sections (4.1.1 to 4.1.3). To start off, it makes sense to consider attacks on “supervisory control and data acquisition (SCADA) systems” (Rivera, 2012, p. 50) because “… SCADA systems [are] used throughout the nation’s critical infrastructure …” (Rivera, 2012, p. 52) – nation here referring to the US – and may be vulnerable to attacks similar to Stuxnet (Rivera, 2012). As a second type of cyberattack, it is important to look at “… Distributed Denial of Service (DDoS) [attacks] … [because] an attack that occurs over an extended time period and prevents access to critical parts of either service or economic infrastructures could be [a] strategic [threat] …” (Rivera, 2012, p. 12). In addition to that, DDoS may also be used as a means to attack SCADA systems (Corfield, 2017). Lastly, ransomware will be considered, in particular because of the “WannaCry [incident which] demonstrated the destructive potential of ransomware …” (Cooper, 2018) by causing “… approximately $4 billion in financial losses” (Cooper, 2018) as well as “… widespread service disruptions at Britain’s National Health Service …” (Cooper, 2018).

The selected types of cyberattacks follow the common pattern of being directed at different types of national infrastructure and should make it possible to cover a spectrum of potential CNA with different capabilities and complexity while staying grounded in reality. Furthermore, it is important to note that, considering the descriptions of the doctrines given by Gaycken and Martellini (2013), the cyberweapons considered as examples in this paper can be used for ‘Assured Disruption’ – which can take the form of ‘Specific Capability’ or ‘General Capability’ – or ‘Silent Erosion’ - as they are limited to CNA as opposed to CNE, i.e. espionage. Therefore, the other doctrines developed by Gaycken and Martellini (2013) which utilize on CNE or other methods entirely – ‘Forced Transparency’, ‘Digital Media Control’, and ‘Attribution Control’ – will not be considered here as it would go beyond the scope of this paper to go into the details of those doctrines and the respective required capabilities.

Lastly, however, this section will take another look at the cost of cyberweapons in general to see if the hypothetical state SM would be able to afford them. According to Hughes and Colarik (2016) small states in particular are faced with “… the escalating costs of military platforms and perceptions that cyber warfare may provide a cheap and effective offensive capability …” (Hughes and Colarik, 2016, p. 19) because

"… the battlespace is open, accessible, nearly anonymous, and with an entry cost that appears affordable to any nation-state" (Hughes and Colarik, 2016, pp. 20–21). Rustici (2011) states that cyberweapons are "… exceedingly cheap …" (Rustici, 2011, p. 34) and give "… small states with minimal defense budgets [the ability] to inflict serious harm on a vastly stronger foe at extreme ranges" (Rustici, 2011, p. 36). In particular, he states that they are much cheaper than conventional weapons that would be able to do similar damage. He goes so far as to state that "[c]yberweapons are a cheap way to build a global strike capability against networked states" (Rustici, 2011, p. 37). A similar argument is made by Gaycken and Martellini (2013), albeit in a less dramatic way. They state that "[o]ffensive cyber capabilities are not as difficult to obtain as most other military capabilities [because] … hacking capabilities are not expensive" (Gaycken and Martellini, 2013, p. 3) and only require "… the brains to design the attack, an intelligence service for reconnaissance and for the deployment of the attack, and testing equipment, depending on the targets they will aim for" (Gaycken and Martellini, 2013, p. 3). Even Valeriano and Maness (2015), who are critical of cyberweapons, agree that they are less expensive. In conclusion, the most common opinion appears to be that "… in many cases cyber weapons represent a cheaper alternative to conventional weapons …" (Maathuis, Pieters and Den Berg, 2016, p. 5). Thus, state SM can be assumed to be theoretically able of developing cyberweapons.

4.1.1 Attacks on SCADA Systems – Iran, Ukraine, and the Future

One of the perhaps most famous cyberweapons is Stuxnet, which was discovered in 2010 when it destroyed "… 1000 of the 5000 centrifuges at Iran's nuclear facilities in Natanz … by mak[ing] the centrifuges [used to enrich uranium for nuclear weapons] spin out of control and ultimately self-destruct" (Keshavarz, 2017b, p. 279). Its discovery marked "… an awareness moment at global level of the existence and utilization of …" (Maathuis, Pieters and Den Berg, 2016, p. 1) cyberweapons. Stuxnet is relevant here because it attacked the machines in the facility "… through supervisory control and data acquisitions (SCADA) systems" (Keshavarz, 2017b, p. 280) which are widely used (Corfield, 2017). According to Rivera (2012), "Stuxnet demonstrates the capability to exploit the vulnerabilities of SCADA systems used throughout … critical infrastructure" (Rivera, 2012, p. 52). Thus, it is necessary to take a closer look at SCADA systems, their vulnerabilities, and known cyberweapons used to attack them. In particular, this section will look at the aforementioned Stuxnet malware as well as

BlackEnergy which was used to attack "… three regional electric power distribution companies of Ukraine … in December 2015" (Khan *et al.*, 2016, p. 3).

SCADA systems, also referred to as "SCADA networks" (Igure, Laughter and Williams, 2006, p. 498), are "… industrial command and control networks …" (Igure, Laughter and Williams, 2006, p. 498) which are widely used across industries because they make it possible that "[p]lant operators … continuously monitor and control many different sections of the plant to ensure its proper operation" (Igure, Laughter and Williams, 2006, p. 498). To do so, the system "… enables data to be collected from remote industrial facilities and instructions sent to control them" (Corfield, 2017, p. 284). Furthermore, SCADA systems "… are the underlying control system of most critical … infrastructures including power, energy, water, transportation, telecommunication …" (Zhu, Joseph and Sastry, 2011, p. 1). In summary, SCADA systems can be found almost anywhere in PN-type states and are of vital importance. Because of that, different authors argue that cyberattacks on SCADA systems could have severe consequences (Zhu, Joseph and Sastry, 2011; Rivera, 2012; Corfield, 2017). To give some examples for potential consequences, Zhu, Joseph and Sastry (2011) state that "[t]hese attacks can disrupt and damage critical infrastructural operations, cause major economic losses, contaminate ecological environment and even more dangerously, claim human lives" (Zhu, Joseph and Sastry, 2011, p. 1).

As stated before, Stuxnet is of particular relevance when talking about cyberattacks on SCADA systems and while different wordings for it can be found in the literature, there appears to be agreement about its high importance. Before Stuxnet, "… it was considered highly unlikely that large scale attacks in the software side of highly specialized applications (such as SCADA) were worth trying or even possible …" (Karnouskos, 2011, p. 1). However, Stuxnet proved the opposite, that such attacks "… are possible and not just theory or movie plotlines" (Falliere, Murchu and Chien, 2011, p. 55). While this paper will not go into detail about the technical details of Stuxnet – they can be found in for example (Falliere, Murchu and Chien, 2011, p. 32) – it is important to address a few points that go beyond the fact that it proved that "[m]alware can affect critical physical infrastructures" (Chen and Abu-Nimeh, 2011, p. 93) and that "… SCADA systems used throughout … critical infrastructure" (Rivera, 2012, p. 52) may be vulnerable (Rivera, 2012).

First of all, "Stuxnet is of … great complexity …" (Falliere, Murchu and Chien, 2011, p. 55) and is classified by Maathuis, Pieters and Den Berg (2016, p. 5) as a “[h]ighly sophisticated” cyberweapon. Similar statements can be found in Chen and Abu-Nimeh (2011), Karnouskos (2011), and Keshavarz (2017b) who all highlight the “… sophistication and complexity …” (Keshavarz, 2017b, p. 281) of Stuxnet. It is argued that this complexity suggests that “… few attackers will be capable of producing a similar threat …” (Falliere, Murchu and Chien, 2011, p. 55) and that it was most likely “… developed by a state rather than an individual or group” (Keshavarz, 2017b, p. 281). Rivera (2012) further states that even though so far few states are capable of developing and deploying such weapons, “… it is unlikely to remain this way in the future” (Rivera, 2012, p. 52). This is especially so when considering that “[e]veryone can download Stuxnet’s source code, modify it and create new cyber weapons” (Maathuis, Pieters and Den Berg, 2016, p. 7).

The second aspect that makes Stuxnet special is that most likely its “… creators had detailed knowledge of its target …” (Chen and Abu-Nimeh, 2011, p. 91) as the malware specifically targets the hardware setup used for the centrifuges in the Iranian facility (Chen and Abu-Nimeh, 2011; Karnouskos, 2011). However, despite this targeting, “Stuxnet’s design and architecture are not domain-specific and … with some modifications it could be tailored as a platform for attacking other systems …” (Karnouskos, 2011, p. 4). Additionally, due to its stealthiness, it may be possible that the malware is not discovered until the damage is already done (Karnouskos, 2011).

The second cyberweapon that will be briefly covered here is BlackEnergy and in particular its deployment to attack “… three regional electric power distribution companies of Ukraine … in December 2015” (Khan *et al.*, 2016, p. 3). Resulting from the attack was a power outage that affected “… 225000 people in western Ukraine …” (Polityuk, Vukmanovic and Jewkes, 2017). It took “… 6+ hours to restore” (Khan *et al.*, 2016, p. 3) power. Furthermore, according to Khan et al. (2016, p. 3) measures were taken by the attacker:

> …[t]o remove attack traces and elongate the blackout period … [by] utiliz[ing] KillDisk malware to wipe/erase several systems and corrupt master boot records in all three companies [as well as installing] … a custom firmware … for serial to Ethernet converters that bricked the devices and prevented technicians from restoring power until converters were bypassed.

Contrary to Stuxnet, which did not have “… the need of any external communication …” (Karnouskos, 2011, p. 3) and contained everything it needed to complete its task (Karnouskos, 2011), a combination of different tools was used to attack the Ukrainian infrastructure (E-ISAC *et al.*, 2016). It is of particular importance that "[t]he outages were caused by the use of control systems and their software through direct interaction by the adversary [while] … other tools …, such as BlackEnergy 3 and KillDisk, were used to enable the attack or delay restoration efforts" (E-ISAC *et al.*, 2016, p. 3). The BlackEnergy 3 malware facilitated the access into the system, which was then essentially remote controlled by the attacker to cause the power outage (E-ISAC *et al.*, 2016; Zetter, 2016). This initial attack was "… combined with amplifying attacks to deny communication infrastructure and future use of their ICSs[[3]] … by destroying equipment and wiping devices …" (E-ISAC et al., 2016, p. 20). Just like Stuxnet, this attack can be classified as "[h]ighly sophisticated" (Maathuis, Pieters and Den Berg, 2016, p. 5). It also shares with Stuxnet that "… it was a first-of-its-kind attack …" (Zetter, 2016). While Stuxnet, according to Keshavarz (2017b), showed for the first time the vulnerability of SCADA system to malware and is sometimes also "… perceive[d] as the first real cyberwarfare weapon" (Chen and Abu-Nimeh, 2011, p. 93), the attack on Ukraine was the first cyberattack on a power grid (Zetter, 2016). Beyond that, it was also the first example for such an attack on "… a nation's critical infrastructure" (E-ISAC *et al.*, 2016, p. 20).

Both of the previously addressed examples demonstrate that the vulnerability of SCADA system does indeed go beyond "… theory or movie plotlines" (Falliere, Murchu and Chien, 2011, p. 55). However, both attacks are described as "[h]ighly sophisticated" (Maathuis, Pieters and Den Berg, 2016, p. 5) and required a great amount of planning and preparation (E-ISAC *et al.*, 2016; Maathuis, Pieters and Den Berg, 2016). With regard to Stuxnet, it can be argued that it "… is of such great complexity … that few attackers will be capable of producing a similar threat …" (Falliere, Murchu and Chien, 2011, p. 55). However, as argued by Rivera (2012), the capability to conduct such attacks may proliferate. Concerning the attack on the Ukrainian power grid, it is important to note that "[n]othing about the attack … was inherently specific to Ukrainian infrastructure" (E-ISAC *et al.*, 2016, p. 20). Furthermore, the result may be more devastating "… in the US, experts say, since many power grid control systems … [lack] manual backup functionality …" (Zetter, 2016) thereby increasing the reliance on SCADA system and making it more difficult to

[3] ICS is the abbreviation for "[i]ndustrial control system …" (*Industrial Control System - Definition - Trend Micro USA*, no date). "There are several types of ICSs, the most common of which are Supervisory Control and Data Acquisition (SCADA) systems, and Distributed Control Systems (DCS)" (*Industrial Control System - Definition - Trend Micro USA*, no date).

recover from an attack (Zetter, 2016). According to Knake (2017) such an attack "… would be extremely difficult but not impossible" (Knake, 2017, p. 1) and may have severe consequences, including power outage "… that could last from days in most places and up to several weeks in others" (Knake, 2017, p. 1) as well as potentially "… widespread injuries and fatalities" (Knake, 2017, p. 3) as a side effect of the power outage. However, it would require a significant investment by the attacker in terms of reconnaissance and development (Knake, 2017).

In summary, "… SCADA systems are vulnerable …" (Corfield, 2017, p. 284) and may become more vulnerable in the future since "[t]he connectivity of SCADA networks with outside networks will continue to grow …" (Igure, Laughter and Williams, 2006, p. 505). The widespread usage of SCADA systems that has been described at the beginning of this section – they are found everywhere in "… critical infrastructure" (Rivera, 2012, p. 52) – implies that such attacks may have far-reaching consequences (e.g. Zhu, Joseph and Sastry, 2011; Rivera, 2012). However, while highly disruptive, both of the examples used in this section, as stated before, required extensive preparation and were very sophisticated (E-ISAC *et al.*, 2016; Maathuis, Pieters and Den Berg, 2016). The same would be true for the hypothetical attack described by Knake (2017), which may take "…months, if not years …" (Knake, 2017, p. 2) to prepare making it perhaps unsuitable as deterrent. The following two sections will take a look at how much of a threat less sophisticated attacks may be to a PN-type state.

4.1.2 Distributed Denial of Service Attacks and Botnets

Distributed Denial of Service (DDoS) attacks were one of "[t]he most populars types of attacks in 2016 …" (Crowther, 2017, p. 77) and "… are perhaps the easiest type of attack to launch …" (Keromytis, 2017b, p. 92). DDoS attacks can take different forms. For instance, a "[t]elephone denial-of-service attack on the call center" (E-ISAC *et al.*, 2016, p. 2) was used in support of the attack on Ukraine described in the previous section to "… deny access to customers reporting outages" (E-ISAC *et al.*, 2016, p. 2). However, "[t]he most common version, a network DDoS, seeks to saturate a target's network links such that there is insufficient bandwidth for legitimate communications" (Keromytis, 2017b, p. 91) with the goal of blocking access to that target for the duration of the attack (Keromytis, 2017b). While there is more than one way to conduct DDoS attacks, "… the most common form of such attacks involves the use of botnets" (Keromytis, 2017b, p. 92), which is essentially "… a group of compromised Internet-connected computers that have been forced to operate on the commands of an unauthorized remote user, usually without the knowledge of the computer's owner" (Henning, 2017, p. 22). Of course, they "… can be used by nation-states … for cyber-warfare operations" (Henning, 2017, p. 24) which is why Botnets and DDoS attacks are addressed here in the context of cybered deterrence.

Though most DDoS attacks only aim at "… remov[ing] access to websites … [and] are more of a nuisance … [,] an attack that occurs over an extended time period and prevents access to critical parts of either service or economic infrastructures could be [a] strategic …" (Rivera, 2012, p. 12) threat to a PN-type state. Like the previous section, this section will again look at two examples, although in this case only one has already happened whereas the other is a theoretical possibility.

The first example "… is the attack on the Domain Name System (DNS) provider Dyn …" (Gerritzen, 2018, p. 6) of "… October 21, 2016 …" (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016, p. 2) that also serves "… as example for the potential of the IoT[[4]] [and DDoS attacks] to be used as a 'weapon of mass disruption' against individual websites or even Internet infrastructure" (Gerritzen, 2018, p. 6). This particular DDoS attack is notable here for two reasons, the first reason being that it "… is the largest and strongest DDoS attack known to date" (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016, p. 4 also cited in Gerritzen, 2018). As a result of the attack "[f]or nearly twelve hours, major internet sites faced either sluggish connection or lack of availability" (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016, p. 4 also cited in Gerritzen, 2018). This is because "Dyn provides critical infrastructure services to major internet sites … such as Twitter, Netflix, or Amazon" (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016, p. 2), which is also the second reason for why it is mentioned here. Since it targeted the DNS – the "… phone book for the Internet" (Gonyea, no date also cited in Gerritzen, 2018) – websites and services relying on Dyn's DNS servers were unavailable (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016).

The second example is related to SCADA systems and describes an additional potential way to attack them. According to Corfield (2017, p. 284), "… SCADA systems are vulnerable to … distributed-denial-of-service (DDOS) attacks". This possibility was analyzed in more detail by Markovic-Petrovic and Stojanovic (2013) through "… comprehensive simulations … assuming a typical IP-based SCADA system

[4] IoT is "short for Internet of things" (*IoT | Definition of IoT in English by Oxford Dictionaries*, no date).

architecture within a power plant” (Markovic-Petrovic and Stojanovic, 2013, p. 2) and comparing the results of the simulation for normal operation with operation when under a DDoS attack. In their model, the SCADA system is connected to the intranet of the organization which is in turn connected to the Internet. This intranet has been compromised by the attacker who created a botnet “… inside the corporate network …” (Markovic-Petrovic and Stojanovic, 2013, p. 3). If the attacker is capable of doing so – their model assumes it has already happened – this then means that “… the attacker can generate traffic similar to legitimate traffic which makes defence mechanisms more difficult” (Markovic-Petrovic and Stojanovic, 2013, p. 3). They conclude that because “[t]here are no safe mechanisms of defence from DDoS attacks … this kind of attacks poses a serious threat to the infrastructure of advanced networks in power generation” (Markovic-Petrovic and Stojanovic, 2013, p. 5). Furthermore, their simulation showed that such an attack could lead “… to a degradation of performances and lack of services of the remote control operating services” (Markovic-Petrovic and Stojanovic, 2013, p. 5) which means that DDoS may indeed be able to disrupt the operation of SCADA systems as stated by Corfield (2017).

In summary, even “… perhaps the easiest type of attack …” (Keromytis, 2017b, p. 92) may cause noticeable issues, as seen in the first example, and even “… wide scale interruptions to … critical infrastructure, including health and safety services” (*October 2016: Black Five Client Advisory, Dyn / DDoS Attack*, 2016, p. 4).

4.1.3 WannaCry and Petya – Ransomware as Cyberweapon?

The last type of cyberattack that will be described in this paper is called *ransomware*. The word is derived from “… the two words *ransom* and *malware*” (Gazet, 2010, p. 77) and describes essentially “… a kind of malware which demands a payment in exchange for stolen functionality” (Gazet, 2010, p. 77). While different types of ransomware exist, “[m]ost widespread ransomwares make an intensive use of file encryption as an extortion mean” (Gazet, 2010, p. 77) which is why this is the only type of ransomware that will be covered here. This malware prevents the target “… from accessing … [their] data using private key encryption until … [they] pay a ransom” (Richardson and North, 2017, p. 10). Ransomware originates from cybercrime where it was used for extortion in the exact method described before (Gazet, 2010; Richardson and North, 2017). It will be covered here for two reasons, which are named WannaCry and Petya/NotPetya.

WannaCry is relevant for the attack that took place on 12 May 2017 which "… demonstrated the destructive potential of ransomware …" (Cooper, 2018) as it "… infect[ed] thousands of computers worldwide within a matter of hours … by exploiting critical vulnerabilities in Windows" (O'Brien, 2017, p. 9). While vulnerabilities and methods to exploit vulnerabilities by themselves are generally nothing special, they are in this case due to their alleged developer. The ransomware made use of "… 'EternalBlue' [which] … had been released … by a group known as the Shadow Brokers, who said the data had been stolen from the Equation group cyber espionage group" (O'Brien, 2017, p. 9). This implies that the exploit that enabled the attack was "… reportedly developed by the U.S. National Security Agency …" (Cooper, 2018) and, following the leak, used by other parties for their own purposes. This exploits enabled the malware to propagate within the intranet of infected organizations which made it particularly dangerous (O'Brien, 2017).

In fact, Cooper (2018) states that WannaCry "… was the most virulent self-spreading malware since 2003 …" (Cooper, 2018); it "… infect[ed] more than 230000 computers systems in 150 countries …" (Cooper, 2018). Furthermore, it caused "… approximately $4 billion in financial losses" (Cooper, 2018) and "… led to widespread service disruptions at Britain's National Health Service, where about 20000 appointments got cancelled as hospitals and clinics were forced offline" (Cooper, 2018). Here it is important to note that the damage caused by WannaCry was below its full potential (O'Brien, 2017). This is because WannaCry was stopped by a security researcher who found and activated a "… kill switch …" (O'Brien, 2017, p. 10) which was embedded into the malware's code (O'Brien, 2017).

However, while WannaCry is notable for the amount of damage it caused, ransomware regularly causes significant losses for companies as a result of outages and disruptions as well as occasionally ransom payments (O'Brien, 2017; Richardson and North, 2017). But this is not the end of the story. According to Richardson and North (2017), "… it seems likely that countries … are looking at ransomware as a potential weapon" (Richardson and North, 2017, p. 15). A similar suspicion is voiced by Cooper (2018), O'Brien (2017) and the Internet Security Threat Report (2018). And this suspicion may have been confirmed already.

While it was originally believed "… to be a WannaCry copycat" (*Internet Security Threat Report*, 2018, p. 39), the Petya/NotPetya attack of 27 June 2018 was in fact “… the most devastating cyberattack since the invention of the internet …” (Greenberg, 2018). The attack caused:

> … more than $10 billion in total damages, according to a White House assessment confirmed to WIRED by former Homeland Security adviser Tom Bossert, who at the time of the attack was President Trumps most senior cybersecurity-focused official (Greenberg, 2018).

Even though there are similarities to WannaCry – it, among other methods, “… also used the EternalBlue exploit” (*Internet Security Threat Report*, 2018, p. 39) to propagate through intranets – Petya/NotPetya was very much different. Whereas WannaCry was clearly ransomware, Petya/NotPetya had no intention of asking for ransom (O’Brien, 2017; *Internet Security Threat Report*, 2018). Instead, it was designed so that “… disks encrypted by Petya/NotPetya could never be recovered” (*Internet Security Threat Report*, 2018, p. 39). While systems infected and encrypted by Petya/NotPetya “… display[ed] an ‘installation key’ which is a randomly generated string” (O’Brien, 2017, pp. 11–12) an – according to O’Brien (2017) and the Internet Security Threat Report (2018) completely unrelated – “… randomly generated … key is … used for disk encryption” (O’Brien, 2017, p. 12). This means that the malware was a “… disk-wiping malware rather than classic ransomware” (O’Brien, 2017, p. 12). It was further equipped with “… a self-propagation mechanism …” (*Internet Security Threat Report*, 2018, p. 39), increasing the reach of the malware far beyond the initially infected systems (*Interest Security Threat Report*, 2018).

In addition to that, contrary to WannaCry, which was not targeted at anything particular according to O’Brien (2017) and the Internet Security Threat Report (2018), this new malware “… was designed to mainly affect organizations in Ukraine” (*Internet Security Threat Report*, 2018, p. 39). It did, however, affect a number of non-Ukrainian organizations, such as “… the world’s largest shipping conglomerate … A.P. Møller-Maersk …” (Greenberg, 2018). Despite them being in the group of “… collateral damage …” (*Internet Security Threat Report*, 2018, p. 39), numerous companies outside of Ukraine suffered hundreds of millions in damage as a consequence of the attack (O’Brien, 2017; Greenberg, 2018). In summary though, the attack “… was highly targeted against Ukraine and deeply disruptive …” (*Internet Security Threat Report*, 2018, p. 39) as well as “… politically motivated …” (O’Brien, 2017, p. 12); it may even be called “… an act of cyberwar …” (Greenberg, 2018).

According to Greenberg (2018), Petya/NotPetya was a ‘cyberweapon’ and a quite effective one at that. But no matter what term one may choose to use for

Petya/NotPetya, it certainly demonstrates the potential of ransomware-like malware to be used as a cyberweapon to cause significant damage and disruption. It also shows that not all cyberweapons have to be complicated as "… ransomware is a cheap and easy form of decoy or disruption" (*Internet Security Threat Report*, 2018, p. 41) and even new exploits can be easily integrated in already existing malware (*Internet Security Threat Report*, 2018). Ransomware is even simpler than "… performing a DDoS attack … [, which] requires a lot more time, effort, and infrastructure" (*Internet Security Threat Report*, 2018, p. 41) and may be used to great effect by a competent attacker.

4.2 Credibility and Escalation

4.2.1 The Credibility Problem of Cyberweapons as Deterrent

The previous sections have shown that the state SM will be able to develop cyberweapons and that those cyberweapons will have – theoretically – the capability to harm a PN-type state. But, as was shown earlier, this is not enough for effective deterrence as it also requires credibility (e.g. Kaufmann, 1954; Long, 2008). Credibility here in particular refers to the "… credibility of capability …" (Long, 2008, p. 11). However, this aspect may be as much of a problem for cybered deterrence as it is important. Valeriano and Maness (2015) covered this problem in detail in the context of cyberdeterrence. They identified that both the nature of cyberweapons and the 'attribution problem' are fundamental problems of cyberdeterrence. While "… attribution of cybered deterrence will not form a problem [because] [i]t will be guaranteed through conventional communication as a deterring force wants to be identified" (Gaycken and Martellini, 2013, p. 4) – this is also why the 'attribution problem' was not addressed in detail before – the nature of cyberweapons may be a problem to cybered deterrence as it does not change whether they are used for cyberdeterrence or cybered deterrence. According to Valeriano and Maness (2015, p. 60), the problem that results from the nature of cyberweapons is a simple causality:

> In order to establish capability, you must demonstrate the capabilities of your weapons; in that process you destroy the advantage, because your cyberweapons are now exposed for all to see. … Deterrence is only in operation when the threat is credible. If the threat cannot be demonstrated, then it cannot be seen as credible, and thus deterrence does not work in cyberspace.

This causality described by Valeriano and Maness (2015) hinges on two assumptions. First, that each cyberweapon "… is a single-use weapon …" (Valeriano and Maness, 2015, p. 59) and second, that it is indeed necessary to use it to create credibility. Both of these points will be analyzed in more detail in the following.

With regard to the first point, this appears to be the consensus in the literature. For instance, even Rustici (2011), who argues very much in favor of the power of cyberweapons, states – using DDoS attacks as a specific example – "… that any attack, even for demonstration purposes, ends up being an irreplicable weapon system" (Rustici, 2011, p. 39). Consequently, "… it is almost impossible to demonstrate cyberpower [because] … any attack results in a near perfect defense within days or at most months …" (Rustici, 2011, p. 38). Similarly, Maathuis, Pieters and Den Berg (2016) identify "[n]o re-use …" (Maathuis, Pieters and Den Berg, 2016, p. 5) as one aspect of cyberweapons but add one important disclaimer, which is that "… if countermeasures are not taken, it is possible to use the same cyber weapon again" (Maathuis, Pieters and Den Berg, 2016, p. 5). Lastly, Hall (2017) concludes that "[c]yberweapons are traditionally thought of as perishable, use-and-lose weapons …" (Hall, 2017, p. 39) but that this does not accurately reflect reality. As already hinted at by Rustici (2011), there is some time required to create and implement the defense (Hall, 2017). Resulting from that "… there is a window of opportunity to re-exploit the vulnerability" (Hall, 2017, p. 39) multiple times.

This can even be shown by taking another brief look at the examples WannaCry and NotPetya. For instance, when WannaCry caused its damage across the world on 12 May 2017, it did so by "… exploiting critical vulnerabilities in Windows, which had been patched two months beforehand by Microsoft" (O'Brien, 2017, p. 9). The very same vulnerabilities where then also used by Petya/NotPetya – "… most devastating cyberattack since the invention of the internet …" (Greenberg, 2018) – on 27 June 2017 (O'Brien, 2017; *Internet Security Threat Report*, 2018). In a perfect world, every IT system would be updated to include the bugfixes against the latest vulnerabilities immediately but, unfortunately, the world is not perfect and in some cases "[p]atching itself may be difficult for key infrastructure systems that much be kept running continuously …" (Hall, 2017, p. 39). Therefore, while cyberweapons are indeed single-use in theory, this may not hold in practice at all times. This should be kept in mind. Still, a state using cybered deterrence has an incentive to not risk making their whole arsenal of cyberweapons unusable by giving their technical details away. Consequently, the first assumption of Valeriano and Maness (2015) holds in practice.

However, this does necessarily mean that the demonstration of capability or deterrence using cyberweapon is impossible. Gaycken and Martellini (2013) address this in more detail and describe how a state may create credibility for their strategy of cybered deterrence. They argue that "… the quality of the military hackers …" (Gaycken and Martellini, 2013, p. 6) is of prime importance and that first and foremost the "… force posture will have to be a proof of an intellectual potential" (Gaycken and Martellini, 2013, p. 6) that is build on two levels. The first level is "… the demonstrat[ion of] efforts and investments in research and development, in personnel or agencies" (Gaycken and Martellini, 2013, p. 6). This would show that there may be the theoretical potential to develop cyberweapons. The second level involves demonstrations of capabilities. Here, the goal would not be to demonstrate every cyberweapon that is developed but to show the "… mastery of cyber warfare" (Gaycken and Martellini, 2013, p. 6) in general. Demonstration can be done either through "… field-test[ing] in the wild or within controlled conditions" (Gaycken and Martellini, 2013, p. 6). In addition to that, the quality and sophistication of the cyberweapons can play into how they are perceived by potential adversaries (Gaycken and Martellini, 2013). In conclusion, Gaycken and Martellini (2013) show that a state may potentially demonstrate capability with compromising their deterrent.

However, there is also another solution to this problem. Rustici (2011) acknowledges that demonstration is impractical which leads to a situation where "… cyber[ed] deterrence is forced to rely almost entirely on a … bluff" (Rustici, 2011, p. 39). In other words, it would be build on uncertainty and fear of potential consequences in an environment where it is difficult to judge what your adversary is actually able to do (Rustici, 2011). While it is possible to argue, correctly so, that "[d]eterrence in many ways requires perfect information" (Valeriano and Maness, 2015, p. 59), uncertainty and fear are important parts of deterrence as well (Long, 2008).

In practice, this means that even though it would be less reliable and may fail, cybered deterrence could still work under conditions that resemble a bluff (Rustici, 2011). This is in particular so if we recall that "… [d]eterrence is the generation of fear" (Long, 2008, p. 6). And whether they are eventually justified or not – Valeriano and Maness (2015) argue that they are not –, there are "… fears associated with cyber technologies …" (Valeriano and Maness, 2015, p. 227). In conclusion, while credibility may not be easy to establish, it is possible for a state to do so.

4.2.2 Flight or Fight – The Risk of Escalation

Even though “[c]redibility is … the linchpin of deterrence, particularly the credibility of threat” (Long, 2008, p. 11) it may – under certain conditions – cause the very event it is trying to deter in the first place (Adamsky, 2013). This is because deterrence only works as a deterrent up to a certain point. (Adamsky, 2013, p. 33) calls this the “… ‘culminating point of deterrence’ [which] is similar to ‘diminishing marginal return’ in economics theory”. The concept of the ‘culminating point of deterrence’ describes how credible threats may lead to escalation instead of deterrence. This is because “[w]hen the ‘culminating point of deterrence’ is crossed a threat becomes more likely to incite the opponent to attack rather than to back down” (Adamsky, 2013, p. 33). Therefore, the ‘culminating point of deterrence’ describes the point of maximum deterrence threat and lowest probably of conflict as a result of deterrence. An increase in the deterrence threat would mean that eventually “… credible threats become of convincing that the adversary feels corned with nothing to lose … and decides to preempt, thus overreacting” (Adamsky, 2013, p. 33). This is a general problem of deterrence threats and therefore it is also relevant to the scenario that is under investigation here. The capability to cause severe damage or disruption in state PN in the hands of a hostile government may cause state SM to cross the ‘culminating point of deterrence’ described by Adamsky (2013), especially since cyberweapons are commonly perceived as a threat (e.g. Rivera, 2012).

However, this is not the only problem of cybered deterrence against interventions that could result in escalation. Considering that, “[a]s these weapons proliferate, it will be increasingly dangerous for … [a PN-type state] to actively shape the international arena through coercive means” (Rustici, 2011, p. 39), state PN may try to disincentive proliferation by not being deterred, an aspect that will be further explained in the following. This possible cause of escalation leads back to the concept of deterrence in general. Deterrence can only work “[i]f leaders view attacking as less risky or costly than any of the alternatives …” (Mazarr, 2018, p. 9) and “… states that initiate aggression … are often responding to situations they perceive as highly dangerous” (Mazarr, 2018, p. 8). In addition to that, Long (2008, p. 9) states the following based on Kahnemann and Tversky (1979, cited in Long, 2008) and Farnham (1994, cited in Long, 2008):

> Experiments in a type of behavioral economics known as prospect theory [show that] … [h]umans as a rule tend to be risk acceptant when facing loss

> and risk averse toward gain. As long as maintaining the status quo is not a clear path to loss, most people will be risk averse in taking steps to upset it. … Of course, two parties may not even agree on what ‘the status quo’ is …

This has an interesting implication for cybered deterrence. It is assumed in the scenario and also described by Rustici (2011) that state PN’s ability to “…shape the international arena through coercive means” (Rustici, 2011, p. 39) is part of their status quo. Thus, it may happen that they are not deterred in order to keep this status quo and to stop future attempts at cybered deterrence, i.e. proliferation, by implementing a cyberdeterrence strategy as for example described by Rivera (2012).

Here it is also important to keep in mind the damage that cyberweapons can cause and to take a brief look at how they may be perceived compared to other weapons. Both Rustici (2011) and Bendiek and Metzger (2015) made this comparison with similar conclusions, placing the value or respectively the level of escalation of powerful cyberattacks above conventional strikes but below nuclear weapons, making cyberweapons quite significant.

Lastly, Gaycken and Martellini (2013) argue that cybered deterrence as chosen method of deterrence itself may cause escalation which is where the ‘attribution problem’ becomes important. Because it is difficult or sometimes even impossible to determine who conducted a specific attack, Gaycken and Martellini (2013, p. 7) arrive at the conclusion that:

> … a demonstrated ability to engage in cyberwarfare in an atmosphere dominated by the problem of attribution will automatically render any actor who demonstrated cyber capabilities into a potential cause of future cyberincidents.

In other words, credibility in the area of cybered deterrence might lead to false accusations with all the consequences that may be attached to them in international relations, which can range from harsh words over sanctions to retaliatory attacks, i.e. escalation to various degrees (Gaycken and Martellini, 2013).

Gaycken and Martellini (2013) further argue that the chosen doctrine may make a difference and assign different escalation potentials to three of the proposed doctrines. They argue that the ‘Targeted Capability’ doctrine has the lowest potential for escalation as the selection of targets is limited. Because of that, only attacks on those targets could lead to false accusations (Gaycken and Martellini, 2013). A higher potential results from the ‘General Capability’ doctrine as the number of potential targets increases (Gaycken and Martellini, 2013). However, both are surpassed by the ‘Silent Erosion’ doctrine. In case of this doctrine, "… the risk of escalation is extraordinarily high" (Gaycken and Martellini, 2013, p. 5) because "[a] 'compensatory hackback spiral' towards massive offensive, state-led hacking could evolve which might eventually cause a real-world escalation as well" (Gaycken and Martellini, 2013, p. 8).

4.3 Can the Powerful, Networked State be Deterred from Intervening?

In the previous sections of this chapter it was shown that cyberweapons are capable of doing significant damage and that they may be used to make a credible deterrence threat. It was also shown that this threat may not necessarily be deterring as it is comes with the potential to escalate the situation. While this is a potential downside of cybered deterrence, it is not the only aspect that needs to be addressed when evaluating cybered deterrence as a "… a form of coercion …" (Long, 2008, p. 8).

The important question remains whether using cyberweapons can change the adversary's behavior in the intended direction which is, as discussed in section 2.1, the fundamental goal of deterrence (e.g. Kaufmann, 1954). To answer this question, the following sections will take a brief look at two aspects. First, the use of cyberweapons in international relations so far and whether they have led to the – sometimes only suspected – intended policy change. Second, the reactions of states when faced with resistance during military interventions, i.e. whether the interventions were continued despite the resistance or if the resistance led to the end of the intervention.

4.3.1 Cyberweapons in International Relations – A Success Story?

When it comes to the effectiveness of cyberweapons in international relations so far, the results may appear to be counter-intuitive. In particular when considering the vast amount of literature on, to name two examples, cyberwar and cyberterrorism as well as the findings of the previous sections, which found cyberweapons to be quite dangerous, it may be surprising that they are not particularly effective as concluded by both Valeriano and Maness (2015) and Iasiello (2013) when they analyzed the impact of different cyberattacks in international relations that are often attributed to state actors.

Both studied "… the 2007 cyber attacks against Estonia …" (Iasiello, 2013, p. 4) as well as the previously addressed Stuxnet. While going into the details of the Estonian attack is beyond the scope of this paper, it is important to note that it is generally believed to be an attempt of Russia to coerce Estonia (Iasiello, 2013; Valeriano and Maness, 2015). Iasiello (2013) additionally considered "… [DDoS] … attacks against the U.S. financial sector" (Iasiello, 2013) in 2012 which are attributed to Iran. Valeriano and Maness (2015) analyzed Shamoon, which is also attributed to Iran and was used "… against Saudi Arabia's national oil and gas firm, Aramco" (Keshavarz, 2017a, p. 154). Based on their analysis of the three different cases, Valeriano and Maness (2015, p. 163) come to the following conclusion:

> Estonia continued to integrate with the West, Iran continued to enrich uranium, and Saudi Arabia continued to support the embargo against Iranian oil. If cyber conflict is used as a tool of the weak and the strong to achieve some sort of political and military ends, it has absolutely failed to this point.

This means that none of the attacks was, despite their technological successes, politically successful as in none of the cases the attacker achieved their goal (Valeriano and Maness, 2015). In addition to that, while all the attacks created fear – which is "… an important result …" (Valeriano and Maness, 2015, p. 162) –, this fear only "… put Estonia closer to Europe …" (Valeriano and Maness, 2015, p. 163) and resulted in increased security in the other two cases (Valeriano and Maness, 2015). This demonstrated failure of cyberweapons may disincentive other actors from attempting similar strategies in the future as "… cyber coercion does not appear to be very effective" (Valeriano and Maness, 2015, p. 162) even when the attacker puts in significant effort, like it was the case with Stuxnet (Valeriano and Maness, 2015).

As stated before, Iasiello (2013) comes to similar conclusions about both Estonia and Stuxnet as well as the overall effectiveness of cyberweapons in international relations. With regard to Estonia he concludes that "… when viewed as an instrument of foreign policy, the DDoS attacks could be considered an unqualified failure that ran the risk of worsening formal relations or escalating into an international incident" (Iasiello, 2013, p. 7). Furthermore, as Valeriano and Maness (2015) had already observed in the case of Stuxnet and Shamoon, Estonia also lead to increased security as it "… incentiviz[ed] NATO's creation of a cyber center of excellence to improve NATO's cyber defense posture" (Iasiello, 2013, p. 7). Like Valeriano and Maness (2015), he further argues that Stuxnet was a failure and "… did not dissuade Iranians" (Iasiello, 2013, p. 10). Instead, both see sanctions against Iran as one of the decisive factors that caused the state's behavior to change. The last example used by Iasiello (2013) are again, like in the case of Estonia, DDoS attacks. In this case the target was "… the U.S. financial sector" (Iasiello, 2013, p. 11). He argues that if Iran is indeed responsible for those attacks, "… the DDoS attacks did not prove to be a viable weapon of influence …" (Iasiello, 2013, p. 14) since they were unsuccessful; the attacks "… made no impact on U.S. plans and intentions towards Iran and its nuclear development, nor did it alter or amend its foreign policy positions” (Iasiello, 2013, p. 14).

Notably, the cases addressed by Valeriano and Maness (2015) and Iasiello (2013) were similar and occasionally also the same as those used as example in this paper. Namely, Stuxnet, DDoS attacks, and ransomware which operated similar to Shamoon even though Shamoon caused significantly less damage as it “… only wiped out the boot sectors in 30000 hard drives [but did not cause] … loss of data or productivity” (Valeriano and Maness, 2015, p. 161).

In conclusion, while none of the addressed cases were instances of attempts at cybered deterrence and not the most damaging attacks that happened so far, the overall lack of success of each attack may imply that cybered deterrence would suffer from similar problems. This is because they “… clearly show … [cyberattacks] to be unsuccessful at influencing decision makers or their courses of action, and therefore [cyberattacks are] not an effective policy tool” (Iasiello, 2013, p. 15).

4.3.2 Military Interventions – Is Resistance Futile?

The previous section covered the past success, or rather lack of success, of cyberattacks in international relations. However, this is only one aspect that is to be addressed with regard to the capability of cybered deterrence against military interventions because, as stated before, it is only a future possibility (Rustici, 2011; Gaycken and Martellini, 2013). The second aspect is the success and failure of military interventions, which is another central aspect of the scenario and plenty of examples exist. Of particular interest here is what causes interventions to be stopped before their completion.

To start off, there is the analysis of Larson (1996) who studied “… the role of casualties in domestic support for U.S. wars and military operations …” (Larson, 1996, p. 99) and reached the conclusion that there is no direct relation based on the cases he studied. Rather, he concludes, it depends on the circumstances of the operation. Specifically, “[t]he historical records suggests that a majority of the American public will be more willing to accept casualties when important interests and principles are at stake” (Larson, 1996, p. 101) as opposed to when this is not the case. In cases where the intervention is perceived as “… lack[ing] either moral force or broadly recognized national interests … even small numbers of casualties may often be sufficient to erode public support for the intervention” (Larson, 1996, p. 100).

This somewhat supports the statement of Rustici (2011) that cybered deterrence may not be effective when "… core interests are at stake …" (Rustici, 2011, p. 40) but very effective when used against "… nonessential security operations" (Rustici, 2011, p. 40), although the public perception of what is a 'nonessential security operations' as opposed to a 'core interest' may differ. In addition to that, the effect may be stronger "… when faced with a catastrophe at home …" (Rustici, 2011, p. 35) than when there are casualties abroad. Rustici (2011) argues that a 'catastrophe at home' reduces support for 'nonessential security operations' when the people perceive it as the cause of the catastrophe (Rustici, 2011). As an example he names "… the Spanish withdrawal from Afghanistan" (Rustici, 2011, p. 35) where a terrorist attack may have caused a change of government to the party supporting the withdrawal. This direct connection between the attacks and the result of the election is supported by Montalvo (2011) whose analysis shows that the result would have been different had the attack not taken place; the party supporting withdrawal would not have won.

The analysis of Larson (1996) is supported by the findings of Burk (1999) who similarly states that casualties and public support are not decisively related. He argues that "… public approval or disapproval … was, in fact, largely determined before casualties occurred" (Burk, 1999, p. 77). This can also be observed in the aforementioned case of Spain as "… the general population never regarded the United States' War on Terror as advancing Spanish national security" (Rustici, 2011, p. 35). Thus, the public already disapproved of the operation before the 'catastrophe at home'.

Using a data set of "… twelve foreign military interventions conducted by the United States or Britain since World War II" (Sullivan, 2008b, p. 120), Sullivan (2008b) confirmed the lack of connection between public support and casualties and additionally found that "… a significant number of individuals that would not have supported *initiating* the use of force at a given set of cost and risk parameters will support *sustaining* an ongoing operation with those parameters" (Sullivan, 2008b, p. 130). Furthermore, this mechanism appears to depend on "… the extent of the state's military commitment" (Sullivan, 2008b, p. 130); higher commitment reduces support for withdrawal (Sullivan, 2008b). This may be a result of "… the costs and psychological impact of withdrawing from military interventions once troops are committed" (Sullivan, 2008b, p. 130).

Another analysis came to the perhaps counter-intuitive result that "[t]here is no relationship between intervention outcomes and relative military capabilities" (Sullivan, 2007, p. 515). Sullivan (2007, p. 519) found that:

> Despite their immense war-fighting capacity, major power states have failed to attain their primary political objective in almost 40 percent of their military operations against weak state and nonstate targets since 1945. In every case, the major power chose to terminate its military intervention short of victory despite the fact that it retained an overwhelming physical capacity to sustain military operations.

The reason for that may be found in the cost of the intervention (Sullivan, 2007). Using the US interventions of the given time-frame as an example, Sullivan (2007) shows that when the US stopped an intervention early, they did so because they "… experienced higher than expected costs and withdrew its troops … despite the fact that its military capacity was at most marginally degraded …" (Sullivan, 2007, p. 519). More broadly stated "[t]he military operations of powerful states are likely to fail only if the state's decision makers initially underestimate the cost …" (Sullivan, 2008a, p. 63).

For the scenario this implies that the capability of the threat to prevent state PN from intervening or to force its early withdrawal depends on the precise circumstances, in particular public perception of the intervention, the commitment of forces, and whether the cost was as expected (Larson, 1996; Burk, 1999; Sullivan, 2007, 2008a, 2008b). However, it is important to keep in mind that it may not work in state SM's favor and that, depending on the circumstances, even a powerful retaliation against state PN may prove futile.

4.4 Cybered Deterrence Game

The previous sections of this chapter have taken a look at the general capability of cyberweapons to inflict damage on a PN-type state, whether the state SM can make their deterrence threat credible, the consequences of that credibility, and lastly whether the state PN can actually be deterred from its military intervention by using cyberweapons. Based on that, this section will define the last details of the cybered deterrence game. Finally, the game will be briefly analyzed before moving on to the simulations.

When it comes to the capability of cyberweapons in general, it has been shown that even those types of attacks described as simple and/or less sophisticated as well as less costly in terms of development and planning – DDoS attacks and ransomware are described as such by Keromytis (2017b) respectively the Internet Security Threat

Report (2018) – can do noticeable and significant damage, as illustrated in particular by WannaCry and Petya/NotPetya (O’Brien, 2017; *Internet Security Threat Report*, 2018). Additionally, cyberweapons appear to have a relatively low entry barrier when it comes to their development (Rustici, 2011; Gaycken and Martellini, 2013; Hughes and Colarik, 2016; Maathuis, Pieters and Den Berg, 2016).

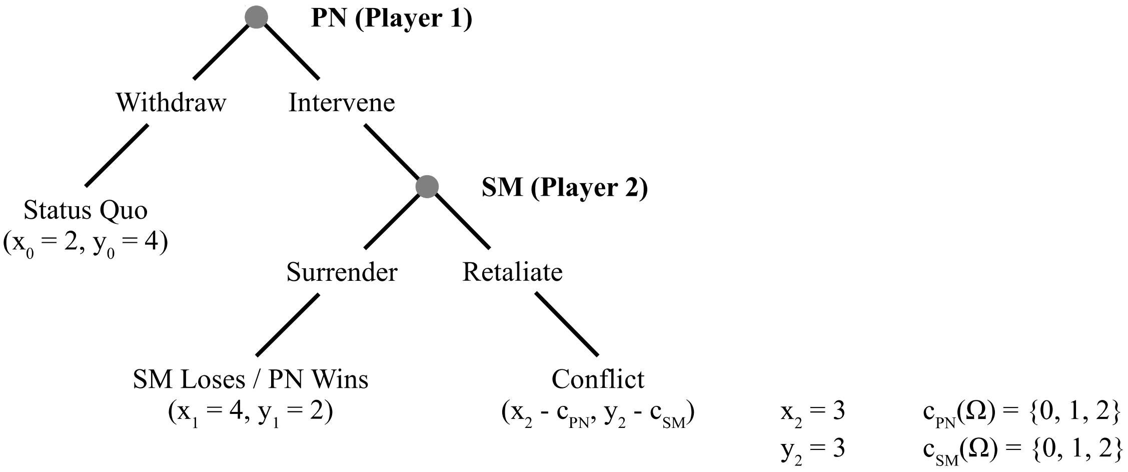
Still, they are not a magical deterrent and come with a number of problems. To start off, credibility – and thus deterrence, considering that “[c]redibility is … the linchpin of deterrence …” (Long, 2008, p. 11) – is difficult, although not impossible as shown by in particular Gaycken and Martellini (2013). However, credibility itself may cause problems and lead to escalation under certain circumstances (Adamsky, 2013; Gaycken and Martellini, 2013). A part of the escalation potential that comes with cyberweapons depends on the cybered deterrence doctrines state SM decides to use (Gaycken and Martellini, 2013). Gaycken and Martellini (2013) give the escalation potential for three of the doctrines, which are, from the lowest to the highest escalation potential, ‘Specific Capability’, ‘General Capability’, and ‘Silent Erosion’. Thus, the simplest doctrine has the highest escalation potential (Gaycken and Martellini, 2013).

When it comes to their capability to induce policy change, the track record of cyberweapons in international relations is not one of political successes but rather of failures (Iasiello, 2013; Valeriano and Maness, 2015). Lastly, as discussed in the previous section, it is debatable whether interventions can even be deterred as it very much depends on the circumstances of the individual intervention (Larson, 1996; Burk, 1999; Sullivan, 2007, 2008a, 2008b). Though, events that occur in the intervening country may change their policy as it was the case in Spain (Montalvo, 2011; Rustici, 2011). Additionally, “… military operations of powerful states are likely to fail only if the state’s decision makers initially underestimate the cost of achieving their objectives” (Sullivan, 2008a, p. 63) and the public may be willing to take the cost if they support the cause of the intervention (Larson, 1996; Burk, 1999; Sullivan, 2008b). If support is there, cost is unlikely to change that (Larson, 1996; Burk, 1999).

In conclusion, when considered in combination, the results of the analysis are unfortunately far from providing a clear, straightforward result on whether the capability (PDT) and credibility (PDT) requirements are fulfilled. Rather, they imply that it depends on the particular context of the intervention in question. This is because the variables c_{PN} and c_{SM} vary with the context. Thus, cybered deterrence may work or not depending on the specific situation surrounding the intervention. This variety of results needs to be considered in the simulations.

For simplicity, the payoffs will be limited to integer values and, as before, the payoff scale of Quackenbush (2011) will be used. The scale allows for payoffs from 1 to 4, 4 being the “best” (Quackenbush, 2011, p. 744) payoff. Since x_2 - c_{PN} < x_1 and y_2 - c_{SM} < y_0 follow from the cost of conflict suggested by Fey and Ramsay (2011, see section 3.4), the maximum payoff for both players for ‘Conflict’ is 3, as the fixed cost of conflict is 1, making it the “next-best” (Quackenbush, 2011, p. 744) outcome. The minimum payoff for both players is 1, in which case it would be the “worst” (Quackenbush, 2011, p. 744) outcome. In order to take into account all possible successes, failures, and reactions, the variables c_{PN} and c_{SM} will be defined as discrete random variables that can take the values 0, 1, or 2. This gives the final ‘Sequential Cybered Deterrence Game’ shown in Figure 5. Table 1 shows the strategic normal form of the game, the ‘Cybered Deterrence Game’ (CDG).

Figure 5: Sequential Cybered Deterrence Game (defined payoffs)



Own illustration, adapted from the "Rudimentary Asymmetric Deterrence Game" (Zagare, 2004, p. 114), the "'Classical' deterrence game" (Quackenbush, 2011, p. 744), the "Simple deterrence game with a credible threat" (Quackenbush, 2011, p. 747), and the "Simple Crisis Game" (Fey and Ramsay, 2011, p. 154).

Table 1: Strategic Normal Form of the Cybered Deterrence Game

		SM (Player 2)	
		Surrender (S)	**Retaliate (R)**
PN (Player 1)	**Withdraw (W)**	2, 4	2, 4
	Intervene (I)	4, 2	x_2 - c_{PN}, y_2 - c_{SM}

Table 2 shows all possible payoff combinations that result from the definitions of x_2, y_2, c_{PN}, and c_{SM}. As stated before, those outcomes should cover all possible successes, failures, and reactions of both players without making the game overly complicated by including many additional variables that would have to be determined. Given that there is a chance of ⅓ for each cost to materialize, the average cost $\bar{c}$ and payoff $\bar{\pi}_{IR}$ are:

$$c = c_{SM}(\Omega) = c_{PN}(\Omega) = \{0,1,2\}$$
$$\bar{c} = \frac{1}{3}(2+1+0) = 1$$
$$x_2 = y_2 = 3$$
$$\bar{\pi}_{IR} = 3 - \bar{c}$$
$$\bar{\pi}_{IR} = 2$$

Lastly, before taking a look at the learning algorithms and simulations based on them, it makes sense to look at what the structure and payoffs of the game itself suggest. This is done by using the game analysis tool Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014). The CDG as it has been set up in Gambit can be found in Appendix C.1. Based on that, Gambit has generated the strategic normal form shown in Table 3, which uses the aforementioned average payoff $\bar{\pi}_{IR}$.

Table 2: Strategic Normal Form of the CDG (Gambit 15.1.1)

		SM (Player 2)	
		Surrender (S)	**Retaliate (R)**
PN (Player 1)	**Withdraw (W)**	2, 4	2, 4
	Intervene (I)	4, 2	2, 2

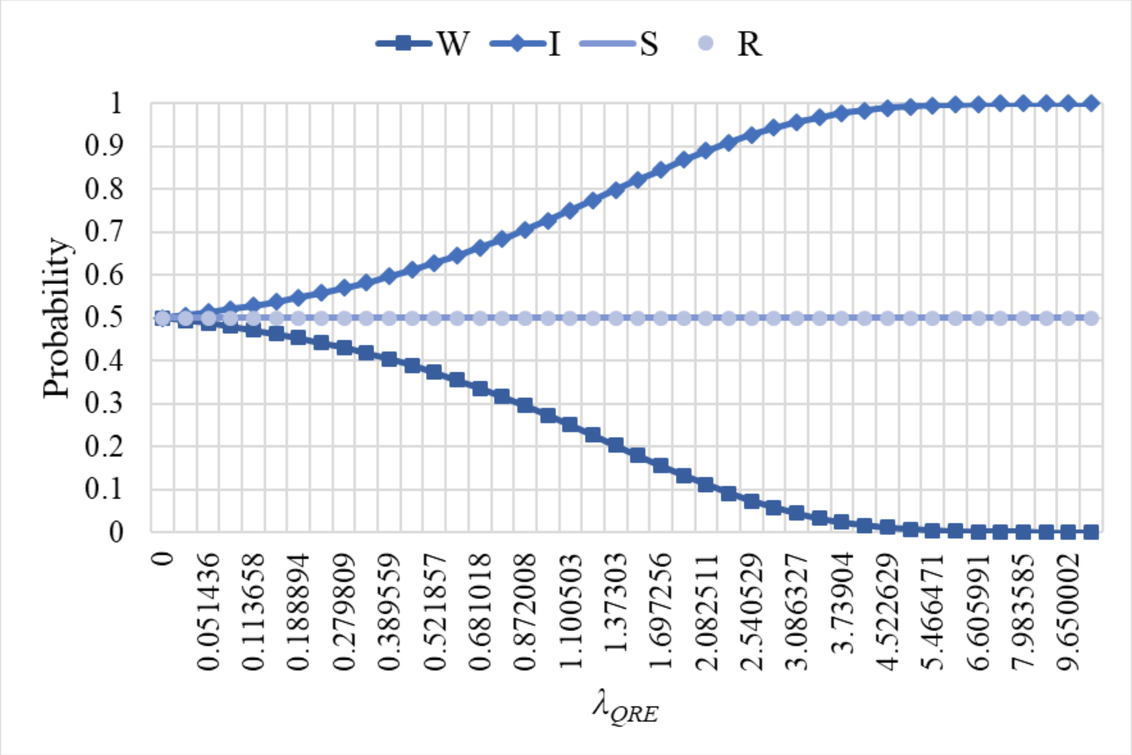
Source: Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014).

Given this game, Gambit identifies three pure strategy Nash Equilibria – “pair[s] of strategies … [that are] a best reply to each other” (Binmore, 2007, p. 18) –, which are WR, IS, and IR. WS is not a Nash Equilibrium as state PN can increase their payoff to IR, thus W is not a best reply to S. Furthermore, it identifies that the strategy I dominates W. Considering that I results in the same payoff as W when played against R and a better payoff when played against S, it fits the definition of a weakly dominant strategy, which is a strategy that “… is … never worse than [another] …, and there is at least one strategy [of the opposing player] that … would make it strictly better” (Binmore, 2007, p. 152).

The last aspect that will be considered here is the “… Quantal Response Equilibrium (QRE) …” (McKelvey and Palfrey, 1995, p. 6) of the CDG. It uses a process that the authors also call “… *learning*” (McKelvey and Palfrey, 1995, p. 8). While it is, as stated by McKelvey and Palfrey (1995), different from the learning algorithms that will be discussed in the following chapter, it still fits into the overall theme of this paper and provides additional information about the game.

The QRE is based on the idea that "… best response functions … [are] probabilistic … rather than deterministic" (McKelvey and Palfrey, 1995, p. 6) and that players do not always select a best response as their strategy. The "… model makes statistical predictions" (McKelvey and Palfrey, 1995, p. 7) to determine with which probability each strategy is selected by the players. Furthermore, it is assumed that "[a]s a player gains experience playing a particular game and makes repeated observations … he/she can be expected to make more precise estimates of the expected payoffs from different strategies" (McKelvey and Palfrey, 1995, p. 8) and therefore make better decisions (McKelvey and Palfrey, 1995). In other words, players learn from previous payoffs. This gaining of experience or learning over time is described by the variable λ_{QRE}. If the variable is equal to 0, the strategy selection is random (McKelvey and Palfrey, 1995). However, "… all limit points of QREs as $\lambda[_{QRE}] \rightarrow \infty$ are Nash equilibria … [and] for almost all games there is a unique selection as $\lambda[_{QRE}] \rightarrow \infty$ … [, which is] the *Limiting Logit Equilibrium* of the game" (McKelvey and Palfrey, 1995, p. 12).

Figure 6: Quantal Response Equilibrium of the CDG (Gambit 15.1.1)



Own illustration, based on the data provided by Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014). A table containing the data calculated by Gambit 15.1.1 can be found in Appendix C.2.

Figure 6 shows the QRE of the CDG as determined by Gambit (McKelvey, McLennan and Turocy, 2014). As can be seen, the behavior of both players starts out random, i.e. both strategies have a probability of 0.5 when λ_{QRE} = 0. Notably, the behavior of state SM remains random and neither S nor R are assigned a probability different from 0.5. This makes sense, given that neither strategy dominates the other; both lead to on average equal payoffs. This is different in the case of state PN as the probability of strategy I increases as $\lambda_{QRE} \rightarrow \infty$ until it reaches 1 whereas W decreases until it reaches 0, i.e. eventually state PN will only play I. In conclusion, this means that the QRE analysis suggests that, as the two players learn, they will eventually reach a mixed strategy Limiting Logit Equilibrium, which is shown in Table 3.

Table 3: Limiting Logic Equilibrium of the CDG (Gambit 15.1.1)

		SM (Player 2)	
	Probabilities (*P*)	$P_{SM}^{S}=0.5$	$P_{SM}^{R}=0.5$
PN (Player 1)	$P_{PN}^{W}=0$	$P_{WS}=0$	$P_{WR}=0$
	$P_{PN}^{I}=1$	$P_{IS}=0.5$	$P_{IR}=0.5$

Source: Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014).

5 Reinforcement Learning and Belief Learning

The ‘Sequential Cybered Deterrence’ and its strategic normal form, the ‘Cybered Deterrence Game’, will be analyzed by applying learning algorithms, also called “… learning model[s] …” (Moffatt, 2016, p. 420). This is done by simulating the behavior of two players who repeatedly play the game and learn (Dhami, 2016; Moffatt, 2016). Specifically, “… *learning* … refer[s] to any change in observed behavior as, ceteris-paribus, players accumulate experience” (Dhami, 2016, p. 1092). Learning in this sense is facilitated by the respective model or algorithm (Dhami, 2016; Moffatt, 2016). For the simulations of the previously defined game, two different learning algorithms will be applied. Those two algorithms are “… reinforcement learning (RL) … [and] … belief learning (BL) …” (Moffatt, 2016, p. 420) which both have particular features, advantages, and disadvantages (Dhami, 2016; Moffatt, 2016). Notably, they can be considered “… special kinds of one learning model” (Camerer and Ho, 1999, p. 828), the “’experience-weighted attraction’ (EWA) model” (Camerer and Ho, 1999, p. 828), which will not be applied here.

Despite their differences, they share a common core concept, which “… is a set of variables known as ‘attractions’” (Moffatt, 2016, p. 423) or “… *propensit*[*ies*]” (Dhami, 2016, p. 1092). Each strategy of each player is assigned one attraction variable (Moffatt, 2016). Based on these attractions is the probability of a strategy to be chosen by the respective player (Moffatt, 2016). However, how the attractions are determined depends on the specific learning algorithm (Moffatt, 2016). The learning algorithms as they are applied in this paper are based on the formulas found in Moffatt (2016), which use the "… notation of Camerer & Ho (1999)" (Moffatt, 2016, p. 423). Furthermore, the method used by Moffatt (2016) to determine the probabilities is that of Camerer & Ho (1999) as well. The formula is given by Moffatt (2016, p. 424, 18.1) as follows:

$$P_i^j(t)=\frac{\exp\left(\lambda A_i^j(t-1)\right)}{\exp\left(\lambda A_i^1(t-1)\right)+\exp\left(\lambda A_i^0(t-1)\right)}$$

It determines the probability *P* of any player *i* to select a strategy *j* in the period *t* by using the previous attractions $A_i^0(t-1)$ and $A_i^1(t-1)$ of the respective player *i*. The numerator uses the attraction for which the probability is determined. All attractions are weighted by the "… sensitivity to attractions …" (Moffatt, 2016, p. 424) *λ* which defines the importance of attractions as a value between 0 and 1 (Moffatt, 2016).

According to Erev and Roth (1998, cited in Moffatt, 2016), "[r]einforcement learning … is … based on the idea that players adjust their strategies in response to *payoffs* received in previous periods" (Moffatt, 2016, p. 424). In practice, this means that there is a direct relation between the payoff received from selecting a particular strategy and the probability of selecting the same strategy in the future (Dhami, 2016); the more successful a strategy is, the more the probability to select it again increases (Dhami, 2016). Moffatt (2016, p. 424, 18.2) shows this mathematically in the formula to calculate attractions under RL:

$$A_i^j(t)=\phi A_i^j(t-1)+I\left(s_i(t)=s_i^j\right)\pi_i\left(s_i^j,s_{-1}(t)\right)$$

This formula gives the attraction *A* to the strategy *j* of player *i* in the period *t* based on three inputs. First, the previous attraction $A_i^j(t-1)$, which is weighted by "[t]he parameter *ϕ* … known as the 'recency' parameter, and indicates the speed at which past payoffs are forgotten …" (Moffatt, 2016, p. 424). Like *λ*, *ϕ* can take a value between 0 and 1 (Moffatt, 2016). The higher *ϕ* is, the more relevant are previous payoffs whereas a value of 0 means "… that only the most recent payoff is remembered …" (Moffatt, 2016, p. 424). Second, the "… indicator function …" (Moffatt, 2016, p. 424) which is designated "*I*(.)" (Moffatt, 2016, p. 424). The indicator function makes sure that the attraction is only updated if the attraction belongs to the strategy $s_i(t)$ selected by the player (Moffatt, 2016). Third, the payoff π of player *i* that results from s_i^j given the opposing player's selection $s_{-i}(t)$ (Moffatt, 2016).

The main aspect of RL is that "… players have minimal rationality and limited information about the game being played" (Dhami, 2016, p. 1092) and thus only use their own strategies, decisions, and respective payoffs as the basis for their attractions and probabilities (Dhami, 2016). This means that they "… do not generally have beliefs about what other players will do" (Camerer and Ho, 1999, p. 828) and also lack information about their opposing players (Dhami, 2016). Because of that, RL is particularly useful for cases where "… the underlying structure of the game is poorly understood by players …" (Dhami, 2016, p. 1093) and less so when the opposite is the case (Dhami, 2016). However, despite this limitation, it generally "… often correctly predicts the direction of learning …" (Dhami, 2016, p. 1102), making it a useful model to consider in this paper. However, one issue of RL is that "… the predicted speed of learning … is quite sluggish relative to the speed of learning that is observed in experiments" (Dhami, 2016, p. 1098).

While RL was developed in the field of psychology, the field of decision and game theory developed a model with a different basic idea (Camerer and Ho, 1999). This model is called "[b]elief learning (BL), also sometimes known as 'weighted fictitious play' …" (Moffatt, 2016, p. 427). The idea behind BL is "… that players … form some belief about what others will do in the future based on past observation" (Camerer and Ho, 1999, p. 828). Instead of *reinforcing* attractions based on the success of the corresponding strategy, "… they tend to choose a best-response, a strategy that maximizes their expected payoffs given the beliefs they formed" (Camerer and Ho, 1999, p. 828). Contrary to RL, the actual success of individual strategies is not relevant in this model as the focus is on which strategy leads to the best results given the strategies previously selected by the adversary (Camerer and Ho, 1999; Dhami, 2016). This way of learning can be simulated with "… the 'weighted fictitious play model'" (Moffatt, 2016, p. 428) (WFPM) for which Moffatt (2016, p. 428, 18.7) gives the following way of calculating attractions:

$$N(t)=\phi N(t-1)+1$$

$$A_i^j=\frac{\phi N(t-1)A_i^j(t-1)+\pi_i\left(s_i^j,s_{-1}(t)\right)}{N(t)}$$

Just like the calculation for RL that was previously explained, the formula determines the attraction *A* to the strategy *j* of player *i* in the period *t*. It furthermore also includes the 'recency parameter' ϕ which has an additional, special importance in the WFPM that will be addressed later (Moffatt, 2016). However, as can be seen when comparing the above calculation with that of RL, there are two notable differences between the algorithms. The first and perhaps most important difference is the absence of the 'indicator function' *I*(.) as "… players adjust their strategies in response to payoffs that *would* have been received under each choice" (Moffatt, 2016, p. 427). In practice this means that both attractions A_i^0 and A_i^1 are updated based on the hypothetical payoff π_i of s_i^0 and s_i^1 given the adversary's strategy $s_{-i}(t)$ (Moffatt, 2016). The second difference is the addition of the "… 'experience' variable *N*(*t*) …" (Moffatt, 2016, p. 428), which "… is a measure of the amount of past experience accumulated at round *t*, measured in 'observation equivalents'" (Moffatt, 2016, p. 429). Here, the special role of ϕ becomes relevant. Moffatt (2016) states that it can turn the WFPM into two special forms of BL, those being "… standard fictitious play (18.5) … [and] the Cournot learning model (18.3)" (Moffatt, 2016, p. 429).

As stated before in the context of RL, ϕ can take a value between 0 and 1 (Moffatt, 2016). When ϕ is set to 0 or 1 in the WFPM, the model resembles the 'Cournot learning model' respectively the 'standard fictitious play' model (Moffatt, 2016). In the case of the 'Cournot learning model', previous outcomes are irrelevant and "… players choose a best response to behavior observed in the previous period" (Moffatt, 2016, p. 427) whereas in the 'standard fictitious play' "… the attraction … is simply the average pay-off in all rounds up to the current round that would have resulted from choosing [the corresponding] strategy …" (Moffatt, 2016, p. 428) because all previous payoffs are considered with the same weight (Moffatt, 2016).

BL is considered here in addition to RL because "… there is a greater degree of rationality" (Dhami, 2016, p. 1106) under BL than under RL and rationality is the main assumption of PDT (Zagare, 2004). This increase in rationality exists because "[p]layers … know the game they are playing, and conditional on their beliefs about the opponent, they play a *best response*" (Dhami, 2016, p. 1106) which is not the case under RL (Dhami, 2016). However, this does not imply that BL is by default the better learning algorithm when compared to RL as shown by Dhami (2016) who compared various analyses dealing with the algorithms' "… empirical performance[s] …" (Dhami, 2016, p. 1094). Dhami (2016, p. 1120) concludes:

> Reinforcement learning appears to fit better than belief-based models in games with mixed strategy equilibrium that have relatively low dimensional strategy spaces (Erev and Roth, 1998; Mookerjee and Sopher, 1994). However, belief-based models make relatively better predictions in coordination games (Battalio et al., 2001; Ho and Wiegelt, 1996).

This means that RL may be overall a better fit for the game in this paper. Nonetheless, both will be simulated the results of all simulations will be considered.

Lastly, is necessary to define a number of variables that influence the behavior of the simulated players (Dhami, 2016; Moffatt, 2016). The first two variables are the 'sensitivity to attractions' λ and the 'recency parameter' ϕ (Moffatt, 2016). In addition to that, "[p]layers are likely to have relevant experience before the start of the game, and this experience is represented by the prior values, $A_i^j(0)$, known as "… 'initial attractions' …" (Moffatt, 2016, p. 424). BL additional requires another value at the start of the simulation, the "… initial experience *N*(0) …" (Moffatt, 2016, p. 429), which defines the importance of the 'initial attractions' (Camerer and Ho, 1999); "[i]f *N*(0) is small the effect of the initial attractions is quickly displaced by experience. If *N*(0) is large then the effect of the initial attractions persists" (Camerer and Ho, 1999, p. 841).

6 Simulations of the SCDG and CDG

This chapter will present and discuss the results of the different simulations. The source code of the respective Python 3 (Python Software Foundation, no date) implementations of the simulations and learning algorithms can be found in Appendix A. Simulations are run of both the SCDG and the CDG with each being simulated under RL and BL. Thus, four models – designated SCDG (RL), SCDG (BL), CDG (RL), and CDG (BL) – have been simulated. All four models are further repeated for three different assumptions for the 'initial attractions'. The following initial attractions are used:

A: $A_{PN}^W(0)=0; A_{PN}^I(0)=0; A_{SM}^S(0)=0; A_{SM}^R(0)=0; N(0)=0$, i.e. no 'initial attractions';

B: $A_{PN}^W(0)=0; A_{PN}^I(0)=4; A_{SM}^S(0)=0; A_{SM}^R(0)=0; N(0)=1$, i.e. state PN has an 'initial attraction towards I of one 'observation equivalent';

C: $A_{PN}^W(0)=0; A_{PN}^I(0)=4; A_{SM}^S(0)=0; A_{SM}^R(0)=4; N(0)=1$, i.e. both have an 'initial attraction' of one 'observation equivalent' towards I or R.

Furthermore, for each set of simulations a series of 'sensitivity to attractions' λ and 'recency parameter' ϕ is used. Each value is iterated in steps of 0.1 between 0 and 1 in such a way that each combination of both values is simulated. This is done to prevent limitation on specific values of ϕ and λ to skew the results of the simulations.

This means that in total there are three simulation super-sets for each model, one for each 'initial attractions' assumption. Within each simulation super-set there is one simulation set for each combination of ϕ and λ, thus there are 121 simulation sets in each of the 12 super-sets. Furthermore, each simulation set contains a number of simulations N depending on whether it is based on the CDG or SCDG.

For the CDG, each set of simulations encompasses a number of simulations $N = 1000$ and a number of periods $M = 100$. In other words, the CDG is repeatedly played 100 times, once per period, which is again repeated for 1000 simulations. For each period the attractions are updated according to the respective algorithm, as described by Dhami (2016) and Moffatt (2016).

The SCDG is simulated with two important differences due to it being a sequential game representing one individual confrontation between the states PN and SM instead of being a regular strategic game. First, whenever the outcome 'Status Quo' or 'SM Loses / PN Wins' is reached, the simulation ends as it means either the withdrawal of state PN from the conflict or the surrender of state SM. Second, while N remains the same, M is reduced significantly and set to $M = 10$. If the conflict continues up to this period, it is assumed that state PN accomplished its objective and thus it is counted as a loss for state SM. The assumption here is that one period may be considered equal to approximately one month and thus the maximum conflict duration is about ten months. This means that the conflict can last longer than "… the median intervention [which] is less than seven months long" (Sullivan and Koch, 2009, p. 713).

Consequently, this also means that the simulations describe different situations. On the one hand, CDG simulations describe the big picture of the international system itself. Each period is one confrontation between the two states that is resolved within the same period with a certain outcome and payoff. Thus, the conflict is repeated M times in one simulation (Moffatt, 2016). SCDG simulations on the other hand are individual confrontations with a certain length until one player withdraws or surrenders. They show what can be expected from an actual conflict between the two states instead of change in the international system, i.e. if retaliation with cyberweapons can force state PN to withdraw or at least increase the overall cost of conflict, i.e. conflict duration.

6.1 Results and Discussion of the CDG Simulations

Table 4: Summary of Average Probabilities in CDG Simulations

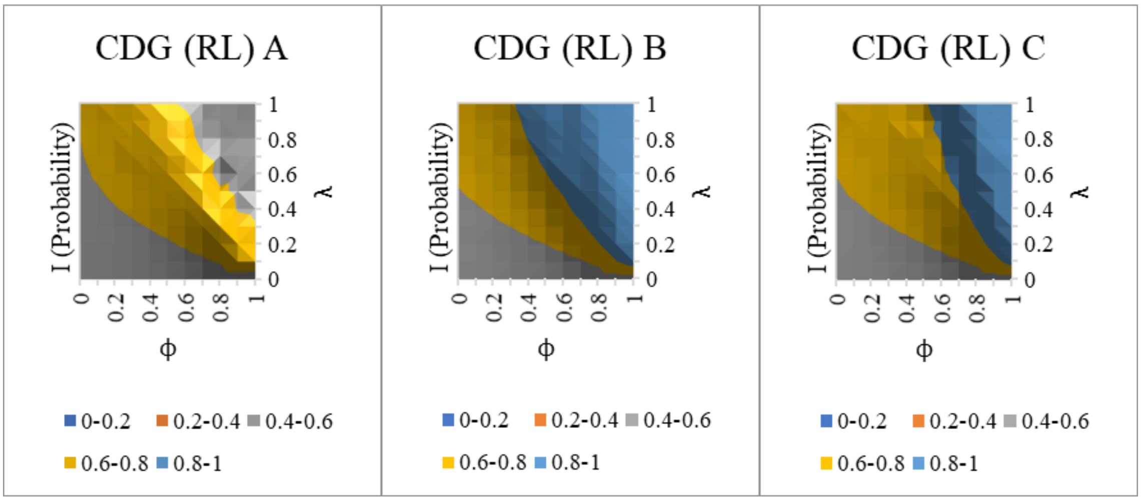
	CDG (RL) A	**CDG (RL) B**	**CDG (RL) C**	**CDG (BL) A**	**CDG (BL) B**	**CDG (BL) C**
$P_{PN}^{\overline{W}}$	0.3964 *SD = 0.672*	0.2541 *SD = 0.1722*	0.2802 *SD = 0.1604*	0.3884 *SD = 0.0667*	0.3810 *SD = 0.0711*	0.3843 *SD = 0.0682*
$P_{PN}^{\overline{I}}$	0.6036 *SD = 0.0672*	0.7459 *SD = 0.1722*	0.7198 *SD = 0.1604*	0.6116 *SD = 0.0667*	0.6190 *SD = 0.0711*	0.6157 *SD = 0.0682*
$P_{SM}^{\overline{S}}$	0.5087 *SD = 0.0161*	0.5192 *SD = 0.0180*	0.3546 *SD = 0.1866*	0.4999 *SD = 0.0003*	0.5000 *SD = 0.0003*	0.4920 *SD = 0.0080*
$P_{SM}^{\overline{R}}$	0.4913 *SD = 0.0161*	0.4808 *SD = 0.0180*	0.6454 *SD = 0.1866*	0.5001 *SD = 0.0003*	0.5000 *SD = 0.0003*	0.5080 *SD = 0.0080*

Table 4 shows the average probability for each strategy to be played by the respective player in all simulation super-sets and provides a broad overview over the behavior of the players. The standard deviations (SD) indicate the deviation of individual simulations from the super-set average, i.e. they show the impact ϕ and λ have on the results of the respective combination of model and assumption. A small SD suggests a small impact of the two variables whereas a large SD suggests the opposite.

Generally, the ‘initial attractions’ have little impact in the CDG (BL) model; the probabilities in A, B, and C are virtually identical and the introduction of the ‘initial attractions’ only causes a small increase of the respective probability. For example, $P_{PN}^{\overline{I}}$ increases slightly in CDG (BL) B and C where the ‘initial attraction’ is introduced. The increase is that small because $N(0) = 1$ is only one ‘observation equivalent’ and thus “… is quickly displaced by experience” (Camerer and Ho, 1999, p. 841). This is different in the CDG (RL) models where even a small increase of the ‘initial attractions’ has a significant effect which results from the mechanism of RL. The ‘initial attraction’ makes it more likely to select that strategy in the beginning and “… strategies are ‘reinforced’ by their previous payoffs …” (Camerer and Ho, 1999, p. 828), further increasing the attraction and probability (Camerer and Ho, 1999).

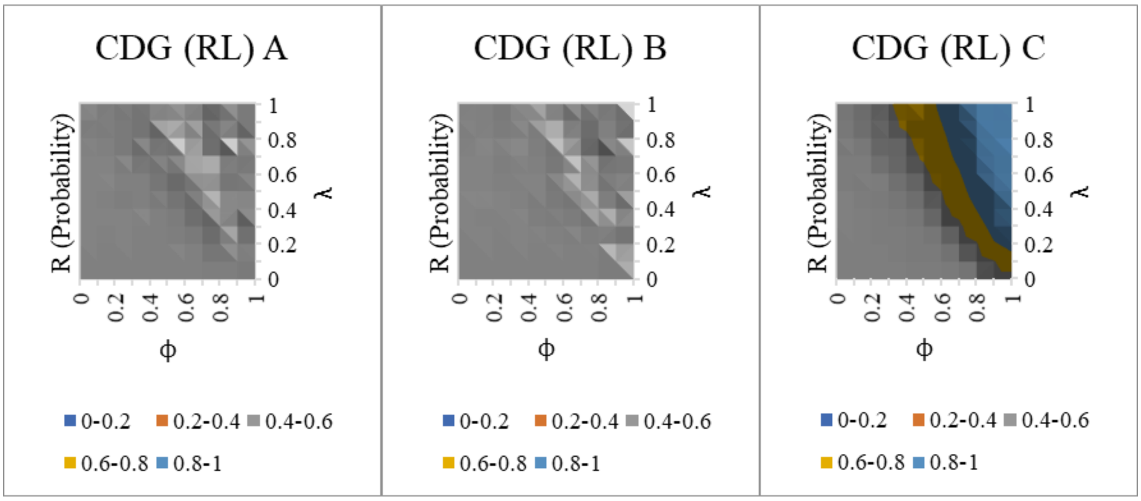
As already suggested by the QRE analysis, state SM’s strategy selection in the CDG will remain essentially random under both RL and BL. The explanation for this is the average cost of conflict $\bar{c}$ which was shown earlier to be 1. This also holds in the simulations. Consequently, the average payoffs of both S and R are identical. However, when the ‘initial attraction’ is introduced in CDG (RL) C, it shifts the behavior of state SM and increases the average probability of playing R to $P_{SM}^{\overline{R}}=0.6454$. Similarly, the 'initial attraction' of state PN towards I, which is introduced in CDG (RL) B and C, increases the super-set average probability of I from $P_{PN}^{\overline{I}}=0.6036$ to $P_{PN}^{\overline{I}}=0.7459$ respectively $P_{PN}^{\overline{I}}=0.7198$. Notably, there is a slight decrease in CDG (RL) C which most likely results from the increased probability of state SM to play R, which in turn reduces the average payoff of playing I for state PN and thus the probability of doing so.

Figure 7: Influence of ϕ and λ on PN in CDG (RL) Simulations



Own illustration. The data for CDG (RL) A and C is provided in Appendix B.1.

Figure 8: Influence of ϕ and λ on SM in CDG (RL) Simulations



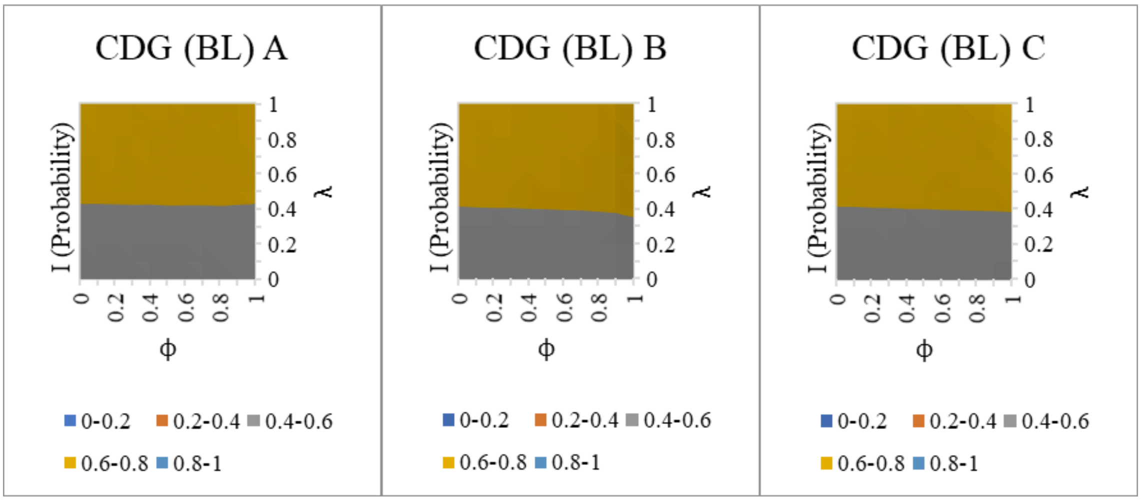
Own illustration. The data for CDG (RL) A and C are provided in Appendix B.1.

As suggested by the SD shown in Table 4, ϕ and λ have a strong impact on the behavior of the players in the RL models, in particular so when 'initial attractions' are introduced. Figure 7 and 8 display the average probabilities $P_{PN}^{\overline{I}}$ respectively $P_{SM}^{\overline{R}}$ for each simulation set with a particular combination of ϕ and λ in the respective super-set. While $P_{SM}^{\overline{R}}$ remains close to 0.5 in A and B, it increasingly shifts towards 1 once the ‘initial attraction’ towards R is introduced. This is in particular so in simulation sets with a high ‘recency parameter’ ϕ, i.e. if past payoffs are important (Moffatt, 2016).

Figure 7 shows a similar pattern of increasing probability for $P_{PN}^{\overline{I}}$ in B and C. Notably, $P_{PN}^{\overline{I}}$ decreases for medium values of ϕ and λ in C compared to B. This may be because the increase of $P_{SM}^{\overline{R}}$ reduces the average payoff of choosing I as it more often results in R instead of S. Thus, $P_{PN}^{\overline{I}}$ decreases slightly in comparison resulting from the reduced average payoff. Contrary to what can be seen in B and C, in A, $P_{PN}^{\overline{I}}$ deviates from random behavior only for medium values of ϕ and λ. This is because for low values, the behavior is random by definition (Moffatt, 2016). For high values, the random (see Appendix B.10) initial “… strategies are ‘reinforced’ by their previous payoffs …” (Camerer and Ho, 1999, p. 828), which, as stated before, increases their future probability (Camerer and Ho, 1999).

In the CDG (BL) model, the SD is small for all $P_{SM}^{\overline{R}} \approx 0.5$ which suggests that it will remain mostly random in all simulation super-sets. While the SD is larger for $P_{PN}^{\overline{I}}$, it remains consistent independent of the ‘initial attractions’. Figure 9 shows the origin of this SD: As λ increases, $P_{PN}^{\overline{I}}$ increases in A, B, and C, i.e. the more important the attractions are, the more likely is state PN to play I instead of choosing randomly.

Figure 9: Influence of ϕ and λ on PN in CDG (BL) Simulations

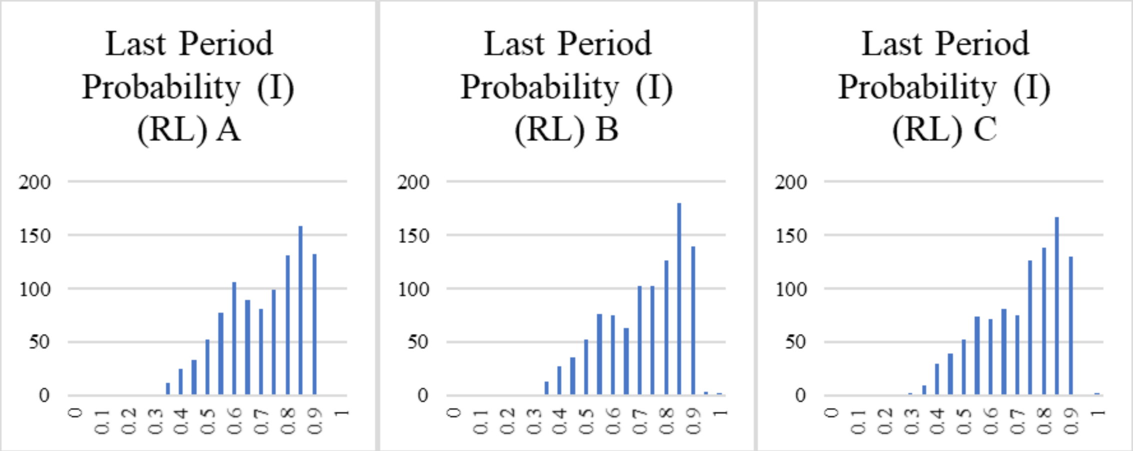


Own illustration. The data for CDG (BL) A is provided in Appendix B.2.

Given that it is reasonable to assume that neither state PN nor state SM will behave completely randomly – i.e. both the ‘recency parameter’ ϕ and the ‘sensitivity to attractions’ λ can be considered to realistically have medium to high values, as this implies less random behavior (Moffatt, 2016) – the CDG model suggests that cybered deterrence will generally not be successful in deterring state PN. This observation is independent of both the learning model and the tested ‘initial attractions’ assumptions. While the simulations suggest that it can work, the average $P_{PN}^{\overline{W}}$ across all simulation super-sets being 0.3474, this is far from implying a reliable deterrent as it suggests that state PN will on average withdraw in only about one third of the cases.

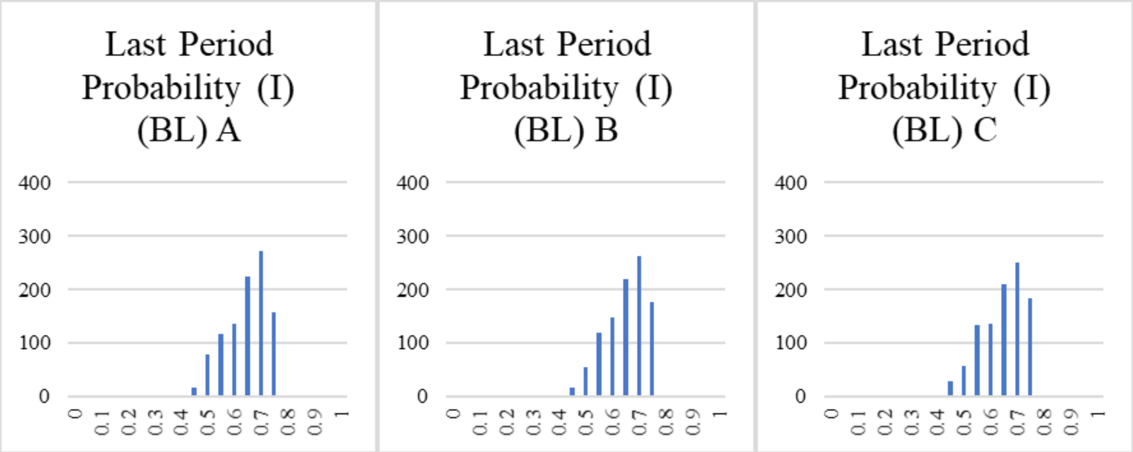
The same pattern is also visible in the final probabilities in individual simulations within one simulation set which can be shown at the example of the simulation sets with ϕ = 0.5 and λ = 0.5 for both RL and BL. Whereas state SM remains more or less indifferent between the two strategies (see Appendix B.3), Figure 10 and 11 show that state PN tends to play I in the last period. This effect is particularly strong in CDG (RL) B and C. As also observed before, the results in the CDG (BL) model are independent of the ‘initial attractions’. Overall, the results do not support cybered deterrence.

Figure 10: Histograms of Final Probabilities of I in CDG (RL; ϕ = 0.5; λ = 0.5)



Own illustration. The data can be found in the GitLab repository belonging to this paper (Gerritzen, 2019).

Figure 11: Histograms of Final Probabilities of I in CDG (BL; ϕ = 0.5; λ = 0.5)



Own illustration. The data can be found in the GitLab repository belonging to this paper (Gerritzen, 2019)

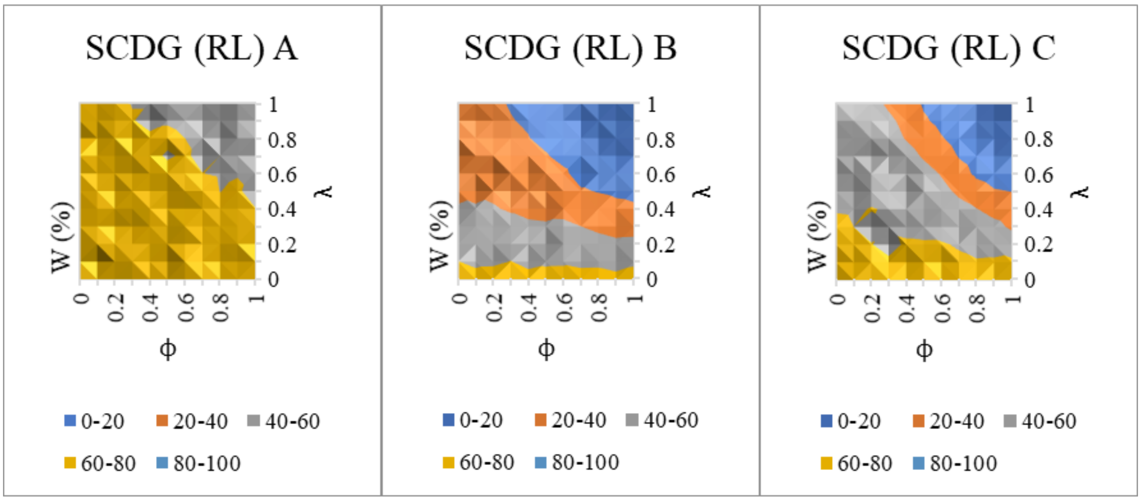
6.2 Results and Discussion of the SCDG Simulations

Table 5: Overview over Results of SCDG Simulations

	SCDG (RL) A	SCDG (RL) B	SCDG (RL) C	SCDG (BL) A	SCDG (BL) B	SCDG (BL) C
Average W (%)	62.60% *SD = 3.99%*	32.62% *SD = 19.06%*	44.09% *SD = 19.86%*	66.39% *SD = 1.60%*	40.21% *SD = 14.37%*	61.33% *SD = 3.25%*
Average S (%)	37.40% *SD = 3.99%*	67.38% *SD = 19.06%*	55.91% *SD = 19.86%*	33.61% *SD = 1.60%*	59.79% *SD = 14.37%*	38.67% *SD = 3.25%*

Table 5 shows an overview over how often W and S are played across the SCDG simulation super-sets. It is important to keep in mind here that in each simulation W or S can only be played once and that they are mutually exclusive. The simulation ends when one of the players decides to play either strategy. At first glance, the results seem to suggest a much more favorable outcome for the cybered deterrence of state SM than in the CDG models. Additionally, the results suggest that in the SCDG simulations the assumption used makes a difference independent of the learning model, although the results and impact of 'initial attractions' differ. Furthermore, the SD suggest that there is again also a significant difference resulting from the 'recency parameter' ϕ and the 'sensitivity to attractions' λ. Therefore, it makes sense to again look at how the average outcome changes depending on ϕ and λ. Figure 12 and 13 show the average W (%) depending on ϕ and λ under RL respectively BL. Since SCDG simulations can only end with either W or S, the results for S mirror those for W, i.e. $S=1-W$. The respective figures are provided in Appendix B.4 but will not be addressed in more detail as it is sufficient to analyze the effect on W; the effect on S is simply the opposite.

Figure 12: Influence of ϕ and λ on Average Occurrence of W in SCDG (RL)

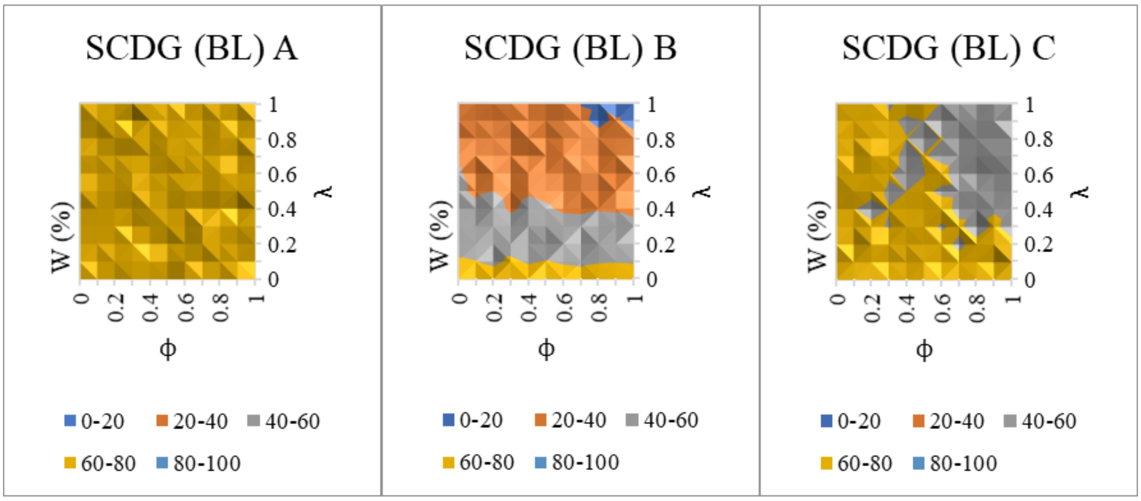


Own illustration. The data for SCDG (RL) A and C is provided in Appendix B.5.

Under RL (see Figure 12), the average W (%) in one simulation set decreases with increasing values of ϕ and λ. This effect can be observed independent of the assumption. It is, however, much more pronounced when the 'initial attraction' towards I is introduced. In SCDG (RL) A, the average W (%) stays around 50% even for high values of ϕ and λ because the initial round is random (see Appendix B.10) and thus the simulation may end instantly if state PN chooses W. When the 'initial attraction' is introduced, I is played relatively more, reducing the overall chance of state PN playing W as ϕ and λ increase. As also observed in the CDG (RL) models, the 'initial attraction' of state SM towards R slightly shifts the distribution of outcomes in state SM's favor.

This is in particular so in the SCDG (BL) models which can be seen in Figure 13. If there are no 'initial attractions', state PN withdraws in approximately 66% of the simulations. When the initial attraction towards I is introduced in B, this changes significantly as state PN withdraws less often, especially for medium and high values of λ. Interestingly, the additional 'initial attraction' of state SM towards R in C causes the results to again resemble more those of A with a slight advantage remaining on the side of state PN which withdraws less for medium values of ϕ and λ.

Figure 13: Influence of ϕ and λ on Average Occurrence of W in SCDG (BL)



Own illustration. The data for SCDG (RL) A, B, and C is provided in Appendix B.6.

In summary, the SCDG simulations suggest that cyberweapons may be effective as deterrent or at forcing early withdrawal given assumption A under RL and BL and assumption C under BL. However, half of the simulation super-sets – SCDG (RL) B, SCDG (RL) C, and SCDG (BL) B – suggest that it will not be effective. This is especially so when considering two aspects. First, as stated before, medium to high values of ϕ and λ can be considered realistic. Second, no 'initial attraction' skews the results as the first period random (see Appendix B.10) does not accurately reflect reality.

The last aspect that will be briefly considered before coming to the conclusion is the duration of the conflict in the SCDG model and the impact of an increased percentage of cyberweapon retaliations on it. An overview can be found in Table 6, which shows that 'initial attractions' increase the average conflict duration under RL and BL, the effect being notably stronger under RL. Again, there is significant variation resulting from ϕ and λ as indicated by the SD. This variation follows the familiar pattern; the average duration increases with ϕ and λ which can be seen in Appendix B.7.

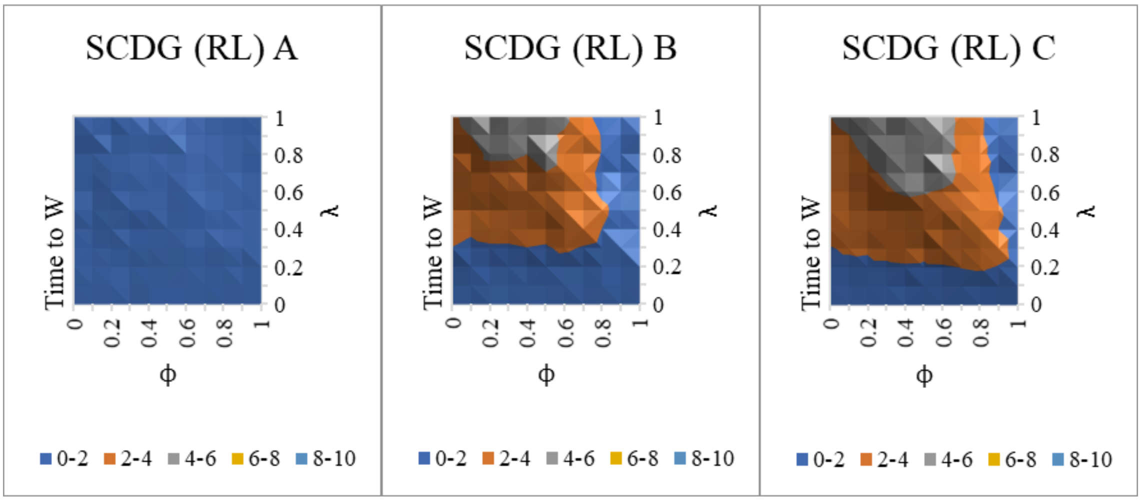
Table 6: Overview over Conflict Duration in SCDG Simulations

	SCDG (RL) A	**SCDG (RL) B**	**SCDG (RL) C**	**SCDG (BL) A**	**SCDG (BL) B**	**SCDG (BL) C**
Average Duration	0.7605 *SD = 0.3835*	1.5593 *SD = 1.0307*	3.3904 *SD = 2.6319*	0.3313 *SD = 0.0225*	0.6050 *SD = 0.1486*	1.1980 *SD = 0.5705*
T (%)	3.75% *SD = 5.11%*	9.36% *SD = 12.51%*	23.64% *SD = 31.51%*	0% *SD = 0%*	0% *SD = 0.91%*	0.01% *SD = 3.61%*
Average to W	1.39 *SD = 0.14*	2.36 *SD = 1.08*	2.73 *SD = 1.18*	1.33 *SD = 0.03*	1.99 *SD = 0.43*	2.24 *SD = 0.59*
Average to S	2.29 *SD = 0.95*	2.49 *SD = 1.10*	4.79 *SD = 2.91*	1.33 *SD = 0.04*	1.43 *SD = 0.09*	2.14 *SD = 0.55*

Furthermore, T (%) gives the average percentage of simulations that ended with the surrender of state SM after 10 periods, i.e. automatic defeat. It can be seen that T (%) is irrelevant in the SCDG (BL) model but reaches up to 23.64% in SCDG (RL) C. However, the SD are large, suggesting a significant variation resulting from different values of ϕ and λ. The influence of ϕ and λ follows the same pattern that can be observed in Figure 12; T (%) increases with ϕ and λ as can be seen in Appendix B.7.

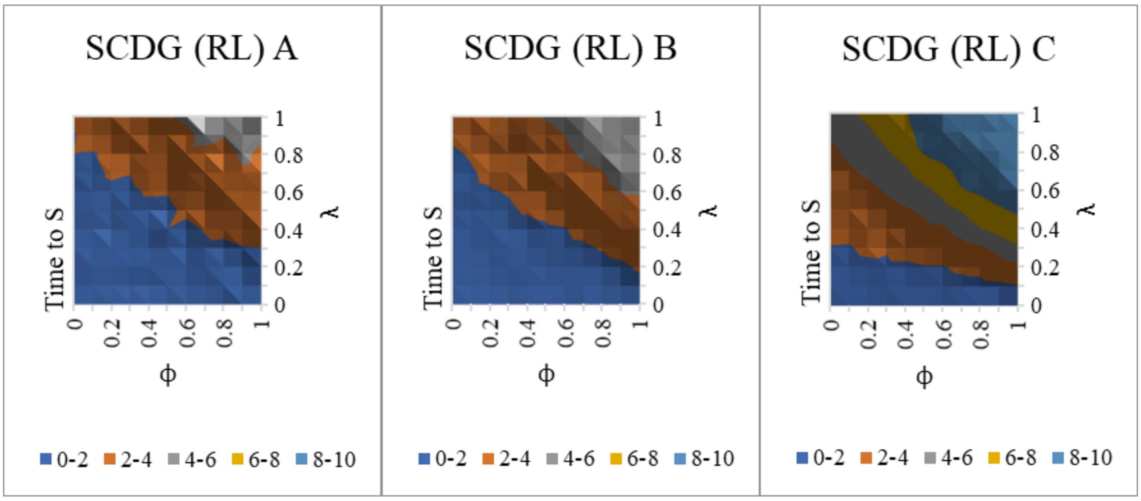
Lastly, it can be observed that the average number of periods (i.e. time) to either W or S increases differently under RL and BL. Under BL there is relatively little variation and most of the variation results – as before – from λ (see Appendix B.8). Additionally, both the average time to W and S increase similarly with the introduction of 'initial attractions'. This is different under RL. Table 6 shows that the average time to S increases more strongly than the average time to W. Figure 15 shows that the average time to S increases with ϕ and λ, as also observed with T (%). The average time to W does not increase as strongly, as shown in Figure 14, and is especially high for medium values of ϕ and high values of λ but remains around 2-4 periods for medium values of both. The results suggest that state PN either withdraws early or eventually wins as state SM either surrenders or is defeated when assumed to have an 'initial attraction' to R.

Figure 14: Influence of ϕ and λ on the Average Time to W in SCDG (RL)



Own illustration. The data for SCDG (RL) C is provided in Appendix B.9.

Figure 15: Influence of ϕ and λ on the Average Time to S in SCDG (RL)



Own illustration. The data for SCDG (RL) C is provided in Appendix B.9.

7 Conclusion

The answer to whether small states can use cyberweapons to deter interventions by powerful, networked state is that a cybered deterrence strategy based on an ‘Assured Disruption’ or ‘Silent Erosion’ doctrine – both developed by (Gaycken and Martellini, 2013) – will be unable to reliably deter interventions. Predictions that cybered deterrence may “… make the world a safer place for corrupt and abusive regimes” (Rustici, 2011, p. 38) are not supported by the results of this paper, which adds to the body of literature that is critical of the predictions made with regard to cyberweapons, a prominent example from that area being the work of Valeriano and Maness (2015).

This does of course not mean that cybered deterrence against interventions is per definition always doomed to fail. It has, however, a number of problems, which make it at worst dangerous to the state using such a strategy and at best unpredictable. Those problems as well as the properties of cyberweapons in general described in particular in chapter four were reflected in the design of the game, which was then analyzed further both as a sequential game, the SCDG, and in its strategic normal form, the CDG. Neither fully supports the concept of cybered deterrence against interventions. This is independent of the learning algorithm used. The CDG in particular suggests that it will not be a viable strategy based on both the simulations and the QRE analysis, which all suggest that it will have a rather low probability of working or that it will not work at all, especially when the simulated players are relatively rational. The SCDG suggests that it may work under some conditions with limited effectiveness but success is far from guaranteed. It further suggests that increased retaliation with cyberweapons may increase the conflict duration and thereby the overall cost of conflict. However, this is no guarantee for deterrence success because the intervening state may expect and accept this cost as necessary (Larson, 1996; Burk, 1999; Sullivan, 2008b).

Of course, cyberweapons are capable of being dangerous as the examples described in the fourth chapter illustrate quite clearly and proper security measures and policies are a necessity. However, one may say that cyberweapons are in general quite overrated, especially when considering their evident lack of political success so far (Iasiello, 2013; Valeriano and Maness, 2015). It remains to be seen whether cyberweapons will eventually live up to the predictions made about them but there appear to be plenty reasons to be skeptical, which this paper has shown at the – of course very specific – example of cybered deterrence against military humanitarian interventions. However, further research is possible and necessary to cover additional aspects or to address specific aspects in more detail than it was possible here. To give three examples, one may, first, analyze the viability of the other cybered deterrence doctrines described by Gaycken and Martellini (2013), which focus on CNE and other methods instead of CNA, in more detail, second, investigate in more depth the aspects covered in this paper to refine the simulations, or third, apply other learning algorithms, such as the EWA of Camerer and Ho (1999), to the same or refined games.

Still, in conclusion, if this paper were to give policy advice, small states should not attempt a strategy of cybered deterrence. As appealing as it might be at first glance, it may have at best a low chance of being successful while being very risky. Conversely, powerful, networked states should be aware that deterrence takes place in their minds and that cyberweapons may not be as scary and powerful as they appear.

References

Adamsky, D. D. (2013) ‘The 1983 Nuclear Crisis – Lessons for Deterrence Theory and Practice’, *Journal of Strategic Studies*, 36(1), pp. 4–41. doi: 10.1080/01402390.2012.732015.

Bendiek, A. and Metzger, T. (2015) ‘Deterrence theory in the cyber-century. Lessons from a state-of-the-art literature review’. Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn.

Binmore, K. G. (2007) *Playing for real: a text on game theory*. Oxford ; New York: Oxford University Press.

Brodie, B. (1959) ‘The Anatomy of Deterrence’, *World Politics*, 11(02), pp. 173–191. doi: 10.2307/2009527.

Burk, J. (1999) ‘Public Support for Peacekeeping in Lebanon and Somalia: Assessing the Casualties Hypothesis’, *Political Science Quarterly*, 114(1), pp. 53–78. doi: 10.2307/2657991.

Camerer, C. and Ho, T.-H. (1999) ‘Experience-weighted Attraction Learning in Normal Form Games’, *Econometrica*, 67(4), pp. 827–874. doi: 10.1111/1468-0262.00054.

Chen, T. M. and Abu-Nimeh, S. (2011) ‘Lessons from Stuxnet’, *Computer*, 44(4), pp. 91–93. doi: 10.1109/MC.2011.115.

Cooper, C. (2018) ‘WannaCry: Lessons Learned 1 Year Later’, *Symantec Blogs*, 16 May. Available at: https://www.symantec.com/blogs/feature-stories/wannacry-lessons-learned-1-year-later (Accessed: 31 October 2018).

Corfield, G. (2017) ‘Supervisory Control and Data Acquisition (SCADA)’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, p. 284.

Crowther, G. A. (2017) ‘Cyber Weapon’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 76–78.

cyberwarfare | Definition of cyberwarfare in English by Oxford Dictionaries (no date) *Oxford Dictionaries | English*. Available at: https://en.oxforddictionaries.com/definition/cyberwarfare (Accessed: 6 January 2019).

cyberweapon | Definition of cyberweapon in English by Oxford Dictionaries (no date) *Oxford Dictionaries | English*. Available at: https://en.oxforddictionaries.com/definition/cyberweapon (Accessed: 6 January 2019).

Dhami, S. S. (2016) ‘Chapter 15: Models of Learning’, in Dhami, S. S., *The foundations of behavioral economic analysis*. First edition. Oxford: Oxford University Press, pp. 1092–1157.

E-ISAC *et al.* (2016) *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Washington, D.C.: SANS ICS. Available at: https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf (Accessed: 7 January 2019).

Falliere, N., Murchu, L. O. and Chien, E. (2011) *W32.Stuxnet Dossier*. Cupertino, California: Symantec Corporation. Available at: https://nsarchive2.gwu.edu//NSAEBB/NSAEBB424/docs/Cyber-044.pdf (Accessed: 5 January 2019).

Fey, M. and Ramsay, K. W. (2011) 'Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict', *American Journal of Political Science*, 55(1), pp. 149–169. doi: 10.1111/j.1540-5907.2010.00486.x.

Gaycken, S. and Martellini, M. (2013) 'Cyber as Deterrent', in Martellini, M. (ed.) *Cyber Security: Deterrence and IT Protection for Critical Infrastructures*. New York: Springer, pp. 1–10.

Gazet, A. (2010) 'Comparative analysis of various ransomware virii', *Journal in Computer Virology*, 6(1), pp. 77–90. doi: 10.1007/s11416-008-0092-2.

George, A. L. and Smoke, R. (1989) 'Deterrence and Foreign Policy', *World Politics*, 41(02), pp. 170–182. doi: 10.2307/2010406.

Gerritzen, C. (2018) 'The Internet of Things: A Weapon of Mass Disruption?' Center for Cyber Security and International Relations Studies. Available at: https://www.cssii.unifi.it/upload/sub/Christopher%20Geritzen%20(1).pdf (Accessed: 12 January 2019).

Gerritzen, C. (2019) *Simulating Cyberweapons as Deterrence Against Humanitarian Interventions Using Game Theory and Learning Algorithms – Simulations and Data*. Available at: https://gitlab.com/gerritzen.christopher/thesis-simulations-interventions-cybered-deterrence (Accessed: 25 February 2019).

Gonyea, C. (no date) 'What is Domain Name Service (DNS)?' Available at: https://dyn.com/blog/dns-why-its-important-how-it-works/ (Accessed: 10 December 2017).

Greenberg, A. (2018) 'The Untold Story of NotPetya, the Most Devastating Cyberattack in History', *Wired*, 22 August. Available at: https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/ (Accessed: 13 January 2019).

Hall, C. G. (2017) *Time Sensitivity in Cyberweapon Reusability*. Naval Postgraduate School.

Henning, L. A. (2017) 'Botnet', in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 22–24.

Hughes, D. and Colarik, A. M. (2016) 'Predicting the Proliferation of Cyber Weapons into Small States', *Joint Forces Quarterly*, 4th Quarter 2016(83), pp. 19–26.

Iasiello, E. (2013) 'Cyber Attack: A Dull Tool to Shape Foreign Policy', in. *2013 5th International Conference on Cyber Conflict*, Tallinn, Estonia: NATO CCD COE Publications.

Iasiello, E. (2014) ‘Is Cyber Deterrence an Illusory Course of Action?’, *Journal of Strategic Security*, 7(1), pp. 54–67. doi: 10.5038/1944-0472.7.1.5.

ICT | Definition of ICT in English by Oxford Dictionaries (no date) *Oxford Dictionaries | English*. Available at: https://en.oxforddictionaries.com/definition/ict (Accessed: 6 January 2019).

Igure, V. M., Laughter, S. A. and Williams, R. D. (2006) ‘Security issues in SCADA networks’, *Computers & Security*, 25(7), pp. 498–506. doi: 10.1016/j.cose.2006.03.001.

Industrial Control System - Definition - Trend Micro USA (no date). Available at: https://www.trendmicro.com/vinfo/us/security/definition/industrial-control-system (Accessed: 11 January 2019).

Internet Security Threat Report (2018). 23. Mountain View, CA: Symantec Corporation. Available at: https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf (Accessed: 13 January 2019).

IoT | Definition of IoT in English by Oxford Dictionaries (no date) *Oxford Dictionaries | English*. Available at: https://en.oxforddictionaries.com/definition/iot (Accessed: 12 January 2019).

Karnouskos, S. (2011) ‘Stuxnet worm impact on industrial cyber-physical system security’, in *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*. *IECON 2011 - 37th Annual Conference of IEEE Industrial Electronics*, Melbourne, Vic, Australia: IEEE, pp. 4490–4494. doi: 10.1109/IECON.2011.6120048.

Kaufmann, W. W. (1954) *The requirements of deterrence*. Princeton, New Jersey: Center of International Studies, Princeton University (Policy Memorandum, 7). Available at: https://babel.hathitrust.org/cgi/pt?id=mdp.39015034327646.

Keromytis, A. D. (2017a) ‘Cyber Deterrence’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 53–55.

Keromytis, A. D. (2017b) ‘Distributed Denial-Of-Service (DDOS) Attack’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 91–93.

Keshavarz, A. (2017a) ‘Iran Cyber Capabilities’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 154–155.

Keshavarz, A. (2017b) ‘Stuxnet’, in Springer, P. J. (ed.) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC, pp. 279–282.

Khan, R. *et al.* (2016) ‘Threat Analysis of BlackEnergy Malware for Synchrophasor based Real-time Control and Monitoring in Smart Grid’, in. *4th International Symposium for ICS & SCADA Cyber Security Research 2016*. doi: 10.14236/ewic/ICS2016.7.

Knake, R. K. (2017) *A Cyberattack on the U.S. Power Grid*. New York, NY: Council on Foreign Relations, Center for Preventive Action.

Larson, E. V. (1996) *Casualties and consensus: the historical role of casualties in domestic support for U.S. military operations*. Santa Monica, CA: RAND.

Lewis, J. A. (2010) *Cross-Domain Deterrence and Credible Threats*. Center for Strategic & International Studies. Available at: http://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/100701_Cross_Domain_Deterrence.pdf (Accessed: 29 September 2018).

Lindsay, J. R. (2015) 'Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack', *Journal of Cybersecurity*, 1(1), pp. 53–67. doi: 10.1093/cybsec/tyv003.

Lindsay, J. R. and Gartzke, E. (2016) 'Cross-Domain Deterrence as a Practical Problem and a Theoretical Concept'. Available at: http://deterrence.ucsd.edu/_files/CDD_Intro_v2.pdf (Accessed: 27 November 2018).

Long, A. G. (2008) *Deterrence: From Cold War to Long War: Lessons from Six Decades of Rand Deterrence Research*. Santa Monica, CA: RAND Corp.

Lupovici, A. (2011) 'Cyber Warfare and Deterrence: Trends and Challenges in Research', *Military and Strategic Affairs*, 3(3), pp. 49–62.

Maathuis, C., Pieters, W. and Den Berg, J. V. (2016) 'Cyber weapons: a profiling framework', in *2016 International Conference on Cyber Conflict (CyCon U.S.). 2016 International Conference on Cyber Conflict (CyCon U.S.)*, Washington, DC, USA: IEEE, pp. 1–8. doi: 10.1109/CYCONUS.2016.7836621.

Markovic-Petrovic, J. D. and Stojanovic, M. D. (2013) 'Analysis of SCADA system vulnerabilities to DDoS attacks', in *2013 11th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services (TELSIKS). TELSIKS 2013 - 2013 11th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services*, Nis, Serbia: IEEE, pp. 591–594. doi: 10.1109/TELSKS.2013.6704448.

Mazarr, M. (2018) *Understanding Deterrence*. RAND Corporation (Perspectives). doi: 10.7249/PE295.

McKelvey, R. D., McLennan, A. M. and Turocy, T. L. (2014) *Gambit: Software Tools for Game Theory*. The Gambit Project. Available at: http://www.gambit-project.org/ (Accessed: 18 December 2018).

McKelvey, R. D. and Palfrey, T. R. (1995) 'Quantal Response Equilibria for Normal Form Games', *Games and Economic Behavior*, 10(1), pp. 6–38. doi: 10.1006/game.1995.1023.

Moffatt, P. G. (2016) 'Chapter 18: Learning Models', in Moffatt, P. G., *Experimetrics: econometrics for experimental economics*. London New York, NY: Macmillan Education, Palgrave, pp. 419–440.

Montalvo, J. G. (2011) ‘Voting after the Bombings: A Natural Experiment on the Effect of Terrorist Attacks on Democratic Elections’, *Review of Economics and Statistics*, 93(4), pp. 1146–1154. doi: 10.1162/REST_a_00115.

Morgan, P. M. (2010) ‘Applicability of Traditional Deterrence Concepts and Theory to the Cyber Realm’, in National Research Council (U.S.) et al., *Proceedings of a workshop on deterring cyberattacks: informing strategies and developing options for U.S. policy*. Washington, D.C.: National Academies Press. Available at: http://public.eblib.com/choice/publicfullrecord.aspx?p=3378670 (Accessed: 6 November 2018).

O’Brien, D. (2017) *Ransomware 2017: An ISTR Special Report*. ISTR Special Report. Mountain View, CA: Symantec Corporation. Available at: https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-ransomware-2017-en.pdf (Accessed: 13 January 2019).

October 2016: Black Five Client Advisory, Dyn / DDoS Attack (2016). Red Five Security, LLC. Available at: http://www.red5security.com/news_media_34_3921121624.pdf (Accessed: 12 October 2017).

Philbin, M. J. (2013) *Cyber Deterrence: An Old Concept in a New Domain*. Strategy Research Project. Carlisle, PA: United States Army War College.

Polityuk, P., Vukmanovic, O. and Jewkes, S. (2017) ‘Ukraine’s power outage was a cyber attack: Ukrenergo’, *Reuters*, 18 January. Available at: https://www.reuters.com/article/us-ukraine-cyber-attack-energy-idUSKBN1521BA (Accessed: 10 January 2019).

Python Software Foundation (no date) *Python*. Python Software Foundation. Available at: https://www.python.org/ (Accessed: 22 January 2019).

Quackenbush, S. L. (2011) ‘Deterrence theory: where do we stand?’, *Review of International Studies*, 37(02), pp. 741–762. doi: 10.1017/S0260210510000896.

Richardson, R. and North, M. (2017) ‘Ransomware: Evolution, Mitigation and Prevention’, *International Management Review*, 13(1), pp. 10–21.

Rid, T. and McBurney, P. (2012) ‘Cyber-Weapons’, *The RUSI Journal*, 157(1), pp. 6–13. doi: 10.1080/03071847.2012.664354.

Rivera, M. (2012) *Deterrence in Cyberspace*. Joint Forces Staff College, Joint Advanced Warfighting School.

Rustici, R. M. (2011) ‘Cyberweapons: Leveling the international playing field’, *Parameters*, 41(3), pp. 32–42.

Springer, P. J. (ed.) (2017) *Encyclopedia of cyber warfare*. Santa Barbara, California: ABC-CLIO, an Imprint of ABC-CLIO, LLC.

Sullivan, P. L. (2007) ‘War Aims and War Outcomes: Why Powerful States Lose Limited Wars’, *Journal of Conflict Resolution*, 51(3), pp. 496–524. doi: 10.1177/0022002707300187.

Sullivan, P. L. (2008a) ‘At What Price Victory? The Effects of Uncertainty on Military Intervention Duration and Outcome’, *Conflict Management and Peace Science*, 25(1), pp. 49–66. doi: 10.1080/07388940701860383.

Sullivan, P. L. (2008b) ‘Sustaining the Fight: A Cross-Sectional Time-Series Analysis of Public Support for Ongoing Military Interventions’, *Conflict Management and Peace Science*, 25(2), pp. 112–135. doi: 10.1080/07388940802007223.

Sullivan, P. L. and Koch, M. T. (2009) ‘Military Intervention by Powerful States, 1945—2003’, *Journal of Peace Research*, 46(5), pp. 707–718. doi: 10.1177/0022343309336796.

‘The GNU General Public License v3.0’ (2007). Free Software Foundation. Available at: https://www.gnu.org/licenses/gpl-3.0.en.html (Accessed: 15 February 2019).

Valeriano, B. and Maness, R. C. (2015) *Cyber war versus cyber realities: cyber conflict in the international system*. Oxford ; New York: Oxford University Press.

Valeriano, B. and Maness, R. C. (2017) ‘International Relations Theory and Cyber Security’. Available at: http://www.brandonvaleriano.com/uploads/8/1/7/3/81735138/international_political_theory_and_cyber_security_-_oxford_handbook_valeriano_and_maness_2018.pdf (Accessed: 10 December 2017).

Zagare, F. C. (2004) ‘Reconciling Rationality with Deterrence: A Re-Examination of the Logical Foundations of Deterrence Theory’, *Journal of Theoretical Politics*, 16(2), pp. 107–141. doi: 10.1177/0951629804041117.

Zetter, K. (2016) ‘Inside the Cunning, Unprecedented Hack of Ukraine’s Power Grid’, *Wired*, 3 March. Available at: https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/ (Accessed: 11 January 2019).

Zhu, B., Joseph, A. and Sastry, S. (2011) ‘A Taxonomy of Cyber Attacks on SCADA Systems’, in *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing. 4th IEEE Int’l Conference on Cyber, Physical and Social Computing (CPSCom)*, Dalian, China: IEEE, pp. 380–388. doi: 10.1109/iThings/CPSCom.2011.34.

Appendix A Simulation Python Code

The Python 3 (Python Software Foundation, no date) code provided in this appendix is licensed under the GPLv3 (‘The GNU General Public License v3.0’, 2007). The code is also available from the GitLab repository belonging to this paper (Gerritzen, 2019).

A.1 Implementation of Learning Algorithms (experimetrics.py)

```
import math
import random

# Function: Strategy selection
def stratselect(prob_i0):
    # Select strategy for player 1
    r = random.uniform(0,1)
    # print(r) # Debug output
    if r <= prob_i0:
        strat_i = 0
    else:
        strat_i = 1
    return strat_i

# Function: Calculation of probability (Moffatt, 2016, p. 424, 18.1)
def probcalc(LAM, attrac_ij, attrac_i0, attrac_i1):
    prob_strat = math.exp(LAM*attrac_ij)/(math.exp(LAM*attrac_i1)+math.exp(LAM*attrac_i0))
    return prob_strat

# Function: Reinforcement learning attraction calculation (Moffatt, 2016, p. 424, 18.2)
def rlcalc(PHI, attrac_prev, payoff):
    attrac_new = PHI*attrac_prev+payoff
    return attrac_new

# Function: Reinforcement learning (Moffatt, 2016, p. 424, 18.2)
def rl(PHI, strat_1, strat_2, payoff_1, payoff_2, attrac_10, attrac_11, attrac_20, attrac_21):
    # Update attractions for player 1
    if strat_1 == 0: # Checks if strategy 1 is selected
        attrac_10 = rlcalc(PHI, attrac_10, payoff_1) # Calls function rlcalc to update the attraction
    elif strat_1 == 1: # Checks if strategy 2 is selected
        attrac_11 = rlcalc(PHI, attrac_11, payoff_1)
    # Update attractions for player 2
    if strat_2 == 0: # Checks if strategy 1 is selected
        attrac_20 = rlcalc(PHI, attrac_20, payoff_2)
    elif strat_2 == 1: # Checks if strategy 2 is selected
        attrac_21 = rlcalc(PHI, attrac_21, payoff_2)
    return attrac_10, attrac_11, attrac_20, attrac_21

# Function: Belief learning attraction calculation (Moffatt, 2016, p. 428, 18.7)
def wfpmcalc(PHI, attrac_prev, payoff, exp_prev, exp_new):
    attrac_new = (PHI*exp_prev*attrac_prev+payoff)/exp_new
    return attrac_new

# Function: Weighted fictitious play model (Moffatt, 2016, p. 428, 18.7)
def wfpm(PHI, payoff_10, payoff_11, payoff_20, payoff_21, attrac_10, attrac_11, attrac_20, attrac_21,
exp_prev):
    # Update experience variable
    exp_new = PHI*exp_prev+1
    # Update attractions for player 1
    attrac_10 = wfpmcalc(PHI, attrac_10, payoff_10, exp_prev, exp_new) # Payoff of strategy 1 given
the choice of player 2
    attrac_11 = wfpmcalc(PHI, attrac_11, payoff_11, exp_prev, exp_new) # Payoff of strategy 2 given
the choice of player 2
    # Update attractions for player 2
    attrac_20 = wfpmcalc(PHI, attrac_20, payoff_20, exp_prev, exp_new) # Payoff of strategy 1 given
the choice of player 1
    attrac_21 = wfpmcalc(PHI, attrac_21, payoff_21, exp_prev, exp_new) # Payoff of strategy 2 given
the choice of player 1
    return attrac_10, attrac_11, attrac_20, attrac_21, exp_new
```

A.2 Configuration File (config.py)

```
# Configuration file for simulations (defaults)

# Enable (True) or disable (False) extended data output
EXTLOG = False
# Number of simulations (can have command line override argv[2])
N = 1000
# Number of periods (can have command line override argv[3])
M_SEQ = 10 # Default number of periods in the sequential game
M_MAT = 100 # Default number of periods in the matrix game
# Sensitivity to attractions (Moffatt, 2016, p. 424)
lam = 1
# Recency parameter (Moffatt, 2016, p. 424)
phi = 1
# If phi = 1, the weighted fictitious play model (wfpm) is equal to standard fictitious play
# If phi = 0, the weighted fictitious play model (wfpm) is equal to Cournot learning model
exp = 0 # Initial experience (wfpm)

# Payoffs
S_SEQ = [2,4,4,2,3,3] # Standard
#S_SEQ = [2,4,4,2,2,2] # Simplified model using average cost instead of random cost
# S_SEQ = [x0,y0,x1,y1,x2,y2]
S_MAT = [2,4,2,4,4,2,3,3] # Standard
#S_MAT = [2,4,2,4,4,2,2,2] # Simplified model using average cost instead of random cost
# S_MAT = [a,b,c,d,e,f,g,h]
# -----------Explanation-----------
# | (a,b) = (2,4) | (c,d) = (2,4) |
# ---------------------------------
# | (e,f) = (4,2) | (g,h) = (3,3) |
# ---------------------------------
# Payoff variation min/max values for player 1 - Cost of conflict of player 1
# Must be 0 if the 'simplified' model is used
C_1_MIN = 0 # Lower boundary of the payoff variation for player 1
C_1_MAX = 2 # Upper boundary of the payoff variation for player 1
# Payoff variation min/max values for player 2 - Cost of conflict of player 2
C_2_MIN = 0 # Lower boundary of the payoff variation for player 2
C_2_MAX = 2 # Upper boundary of the payoff variation for player 2

# Initial attraction values, usually set to 0
attrac_10 = 0 # Player 1: Initial attraction to strategy 1 (w)
attrac_11 = 0 # Player 1: initial attraction to strategy 2 (i)
attrac_20 = 0 # Player 2: Initial attraction to strategy 1 (s)
attrac_21 = 0 # Player 2: Initial attraction to strategy 2 (r)
```

A.3 Sequential Cybered Deterrence Game (seqgame.py)

```
import math
import random
import csv
import experimetrics
from config import EXTLOG

# Function: Simulation for 'Sequential Cybered Deterrence Game'
def scdg(S, N, M, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM):
    # Create output files
    if EXTLOG == True:
        logFile = open('output/' + TIME + '-' + LM + '/lam' + str(lam) + '/' + 'log-' + TIME + '-phi'
+ str(phi) + '-lam' + str(lam) + '.txt', 'w') # Create log file
        dataFile = open('output/' + TIME + '-' + LM + '/lam' + str(lam) + '/' + 'data-' + TIME + '-
phi' + str(phi) + '-lam' + str(lam) + '.csv', 'w', newline='') # Create data file
        dataWriter = csv.writer(dataFile) # Create data file writer
        dataWriter.writerow(['Simulation', 'Period', 'attrac_10 (w)', 'attrac_11 (i)', 'attrac_20
(s)', 'attrac_21 (r)',
                            'prob_10 (w)', 'prob_11 (i)', 'prob_20 (s)', 'prob_21 (r)', 'strat_1',
'strat_1_id', 'strat_2', 'strat_2_id',
                            'payoff_1', 'payoff_2', 'cost_1', 'cost_2'])

    # Initialize variables
    sum_time_to_withdrawal = 0
    sum_time_to_surrender = 0
    sum_payoff_1 = 0
    sum_payoff_2 = 0
    sum_cost_1 = 0
    sum_cost_2 = 0
    sum_prob_10 = 0
    sum_prob_11 = 0
    sum_prob_20 = 0
    sum_prob_21 = 0
    n_withdrawal = 0
    n_surrender = 0
    n_conflict = 0
    n_timeout = 0
    n_periods = 0
    # Simulations
    for sim in range(0,N): # Loop that runs N simulations

        from config import exp # Only used for wfpm (experience variable)
        from config import attrac_10, attrac_11, attrac_20, attrac_21 # Resets the attraction values
at the beginning of every simulation to the starting values
        end_simulation = False # Initialize end simulation trigger

        for period in range(0,M): # Loop that runs M periods within every simulation

            n_periods = n_periods+1 # Count total number of periods within one set of simulations

            # Update probabilities
            prob_10 = experimetrics.probcalc(lam, attrac_10, attrac_10, attrac_11) # Player 1: Update
probability of strategy 1 (w)
            prob_11 = experimetrics.probcalc(lam, attrac_11, attrac_10, attrac_11) # Player 1: Update
probability of strategy 2 (i)
            prob_20 = experimetrics.probcalc(lam, attrac_20, attrac_20, attrac_21) # Player 2: Update
probability of strategy 1 (r)
            prob_21 = experimetrics.probcalc(lam, attrac_21, attrac_20, attrac_21) # Player 2: Update
probability of strategy 2 (s)
            # Save sum of probabilities
            sum_prob_10 = sum_prob_10+prob_10
            sum_prob_11 = sum_prob_11+prob_11
            sum_prob_20 = sum_prob_20+prob_20
            sum_prob_21 = sum_prob_21+prob_21

            # Select strategy of player 1
            strat_1 = experimetrics.stratselect(prob_10)

            # Player 1 decides not to intervene
            if strat_1 == 0: # Determine payoffs for case player 1 does not intervene
                payoff_1 = S[0]
                payoff_2 = S[1]
                sum_time_to_withdrawal = sum_time_to_withdrawal+period+1 # +1 to count the initial
period as first instead of null period
                n_withdrawal = n_withdrawal+1 # Count simulations that end with withdrawal of player
1
                # Print output
                if EXTLOG == True:
                    print('Simulation ' + str(sim) + '/' + str(period) + ': Player 1 withdraws' + '\
n' +
                          '→ Player 2 wins: Successful deterrence', file=logFile) # Log output
                    dataWriter.writerow([sim, period, attrac_10, attrac_11, attrac_20, attrac_21,
prob_10, prob_11, prob_20, prob_21, strat_1, 'w', '', '', payoff_1, payoff_2])
                # Trigger end simulation
                end_simulation = True

            # Player 1 decides to intervene
            elif strat_1 == 1: # Determine payoffs for case player 1 intervenes based on decision of
player 2
                if EXTLOG == True:
                    print('Simulation ' + str(sim) + '/' + str(period) + ': Player 1 intervenes',
file=logFile) # Log output

                if period == M-1:
                    strat_2 = 0
                    n_timeout = n_timeout+1 # Count simulations that end with surrender of player 2
after timeout
                    if EXTLOG == True:
                        print('Player 2 has lost the war', file=logFile)
                else:
                    strat_2 = experimetrics.stratselect(prob_20)

                # Player 2 decides to surrender
                if strat_2 == 0:
                    payoff_1 = S[2]
                    payoff_2 = S[3]
                    sum_time_to_surrender = sum_time_to_surrender+period+1
                    n_surrender = n_surrender+1 # Count simulations that end with surrender of player
2
                    # Print output
                    if EXTLOG == True:
                        dataWriter.writerow([sim, period, attrac_10, attrac_11, attrac_20, attrac_21,
prob_10, prob_11, prob_20, prob_21, strat_1, 'i', strat_2, 's', payoff_1, payoff_2])
                        print('Simulation ' + str(sim) + '/' + str(period) + ': Player 2 surrenders'
+ '\n' +
                              '→ Player 1 wins: Successful intervention', file=logFile) # Log output
                    # Trigger end simulation
                    end_simulation = True

                # Player 2 decides to retaliate
                elif strat_2 == 1:
                    cost_1 = random.randint(C_1_MIN,C_1_MAX) # Cost of conflict player 1
                    cost_2 = random.randint(C_2_MIN,C_2_MAX) # Cost of conflict player 2
                    payoff_1 = S[4]-cost_1
                    payoff_2 = S[5]-cost_2
                    n_conflict = n_conflict+1 # Count periods of conflict
                    sum_cost_1 = sum_cost_1+cost_1
                    sum_cost_2 = sum_cost_2+cost_2
                    # Print output
                    if EXTLOG == True:
                        dataWriter.writerow([sim, period, attrac_10, attrac_11, attrac_20, attrac_21,
prob_10, prob_11, prob_20, prob_21, strat_1, 'i', strat_2, 'r', payoff_1, payoff_2, cost_1, cost_2])
                        print('Simulation ' + str(sim) + '/' + str(period) + ': Player 2 retaliates',
file=logFile) # Log output
                        print('PAYOFF VARIATION: cost_1 = ' + str(cost_1) + '; cost_2 = ' +
str(cost_2), file=logFile) # Log output

            # Update payoff sum
            sum_payoff_1 = sum_payoff_1+payoff_1
            sum_payoff_2 = sum_payoff_2+payoff_2

            # End simulation if triggered
            if end_simulation == True:
                break
            else:
                if LM == 'rl':
                    # Update attractions
                    attrac_10, attrac_11, attrac_20, attrac_21 = experimetrics.rl(phi, strat_1,
strat_2, payoff_1, payoff_2, attrac_10, attrac_11, attrac_20, attrac_21)
                elif LM == 'wfpm':
                    # Payoffs of conflict and hypothetical payoffs of alternatives
                    payoff_10 = S[0]
                    payoff_11 = payoff_1
                    payoff_20 = S[3]
                    payoff_21 = payoff_2
                    # Update attractions
                    attrac_10, attrac_11, attrac_20, attrac_21, exp = experimetrics.wfpm(phi,
payoff_10, payoff_11, payoff_20, payoff_21, attrac_10, attrac_11, attrac_20, attrac_21, exp)

    if EXTLOG == True:
        logFile.close()
        dataFile.close()

    average_time_to_withdrawal = sum_time_to_withdrawal/n_withdrawal
    average_time_to_surrender = sum_time_to_surrender/n_surrender

    return n_periods, n_conflict, average_time_to_withdrawal, n_withdrawal,
average_time_to_surrender, n_surrender, n_timeout, sum_payoff_1, sum_payoff_2, sum_cost_1,
sum_cost_2, sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21
```

A.4 Simulation Series SCDG (simulations_seq_phi-lam-series.py)

```
import math
import csv
import time
import os
import seqgame
import sys
from tqdm import trange # Import tqdm progress bar
from config import S_SEQ, N, M_SEQ, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, EXTLOG

TIME = time.strftime("D%Y%m%dT%H%M%S") # Date/time of simulation
LM = str(sys.argv[1]) # Get command line argument 1
print('Learning model: ' + str(sys.argv[1])) # Print used learning model to console

# Create file for summary table
os.makedirs('output/' + TIME + '-' + LM)
summaryTableFile = open('output/' + TIME + '-' + LM + '/' + 'summaryTable-' + TIME + '.csv', 'w',
newline='') # Create summary table file
summaryTableWriter = csv.writer(summaryTableFile) # Create data file writer
summaryTableWriter.writerow(['lambda', 'phi', 'n_periods', 'n_conflict',
                             'average time to withdrawal', 'n_withdrawal', 'percentage successful
deterrence',
                             'average time to surrender', 'n_surrender', 'percentage successful
intervention',
                             'n_timeout', 'percentage timeout',
                             'sum_payoff_1', 'sum_payoff_2', 'average payoff 1', 'average payoff
2','average cost 1', 'average cost 2',
                             'average prob_10 (w)', 'average prob_11 (i)', 'average prob_20 (s)',
'average prob_21 (r)'])

# Run simulations for lambda between 0 and 1
for lam in trange(0,11,desc='Lambda'):
    lam=lam/10 # range() function only supports integers, convert integer to respective float

    if EXTLOG == True:
        os.makedirs('output/' + TIME + '-' + LM + '/lam' +str(lam)) # Create sub-directories for each
lambda

    # Run simulations for phi between 0 and 1
    for phi in trange(0,11,desc='Phi   '):
        phi=phi/10 # range() function only supports integers, convert integer to respective float
        (n_periods, n_conflict, average_time_to_withdrawal, n_withdrawal, average_time_to_surrender,
         n_surrender, n_timeout, sum_payoff_1, sum_payoff_2, sum_cost_1, sum_cost_2,
         sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21) = seqgame.scdg(S_SEQ, N, M_SEQ, C_1_MIN,
C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM)

        # Write summary to summary table (1 line for each lam/phi combination)
        summaryTableWriter.writerow([lam, phi, n_periods, n_conflict,
                                     average_time_to_withdrawal, n_withdrawal, n_withdrawal/N*100,
                                     average_time_to_surrender, n_surrender, n_surrender/N*100,
                                     n_timeout, n_timeout/N*100,
                                     sum_payoff_1, sum_payoff_2, sum_payoff_1/n_periods,
sum_payoff_2/n_periods, sum_cost_1/n_conflict, sum_cost_2/n_conflict,
                                     sum_prob_10/n_periods, sum_prob_11/n_periods,
sum_prob_20/n_periods, sum_prob_21/n_periods])

# Close summary table file and print statement of completion
summaryTableFile.close()
print('\nCompleted ' + str(N) + ' simulations for every phi and lambda between 0 and 1 iterated at
0.1')
```

A.5 Simulation SCDG (simulations_seq_phi-lam-static.py)

```
import math
import csv
import time
import os
import seqgame
import sys
from config import S_SEQ, N, M_SEQ, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, EXTLOG

TIME = time.strftime("D%Y%m%dT%H%M%S") # Date/time of simulation
LM = str(sys.argv[1]) # Get command line argument 1
print('Learning model: ' + str(sys.argv[1])) # Print used learning model to console

if EXTLOG == True:
    os.makedirs('output/' + TIME + '-' + LM + '/lam' +str(lam)) # Create sub-directory for lambda

(n_periods, n_conflict, average_time_to_withdrawal, n_withdrawal, average_time_to_surrender,
 n_surrender, n_timeout, sum_payoff_1, sum_payoff_2, sum_cost_1, sum_cost_2,
 sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21) = seqgame.scdg(S_SEQ, N, M_SEQ, C_1_MIN,
C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM)
```

A.6 2x2 Matrix Game (matrixgame.py)

```
import math
import random
import csv
import experimetrics
from config import EXTLOG

# Function: Simulation for 2x2 matrix game
def twobytwo(S, N, M, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM):
    # Create output files
    if EXTLOG == True:
        dataFile = open('output/' + TIME + '-' + LM + '/lam' + str(lam) + '/' + 'data-' + TIME + '-
phi' + str(phi) + '-lam' + str(lam) + '.csv', 'w', newline='') # Create data file
        dataWriter = csv.writer(dataFile) # Create data file writer
        dataWriter.writerow(['Simulation', 'Period', 'attrac_10', 'attrac_11', 'attrac_20',
'attrac_21',
                            'prob_10', 'prob_11', 'prob_20', 'prob_21', 'strat_1', 'strat_2',
                            'payoff_1', 'payoff_2', 'cost_1', 'cost_2'])

    # Initialize variables
    sum_payoff_1 = 0
    sum_payoff_2 = 0
    sum_cost_1 = 0
    sum_cost_2 = 0
    sum_prob_10 = 0
    sum_prob_11 = 0
    sum_prob_20 = 0
    sum_prob_21 = 0
    n_00 = 0
    n_01 = 0
    n_10 = 0
    n_11 = 0
    n_periods = 0

    # Simulations
    for sim in range(0,N): # Loop that runs N simulations

        from config import exp # Only used for wfpm (experience variable)
        from config import attrac_10, attrac_11, attrac_20, attrac_21 # Resets the attraction values
at the beginning of every simulation to the starting values

        for period in range(0,M): # Loop that runs M periods within every simulation

```

```
            n_periods = n_periods+1

            # Update probabilities
            prob_10 = experimetrics.probcalc(lam, attrac_10, attrac_10, attrac_11) # Player 1: Update
probability of strategy 1
            prob_11 = experimetrics.probcalc(lam, attrac_11, attrac_10, attrac_11) # Player 1: Update
probability of strategy 2
            prob_20 = experimetrics.probcalc(lam, attrac_20, attrac_20, attrac_21) # Player 2: Update
probability of strategy 1
            prob_21 = experimetrics.probcalc(lam, attrac_21, attrac_20, attrac_21) # Player 2: Update
probability of strategy 2
            # Save sum of probabilities
            sum_prob_10 = sum_prob_10+prob_10
            sum_prob_11 = sum_prob_11+prob_11
            sum_prob_20 = sum_prob_20+prob_20
            sum_prob_21 = sum_prob_21+prob_21
            # Determine cost variables
            cost_1 = random.randint(C_1_MIN, C_1_MAX)
            cost_2 = random.randint(C_2_MIN, C_2_MAX)
            sum_cost_1 = sum_cost_1+cost_1
            sum_cost_2 = sum_cost_2+cost_2

            # Select strategy of player 1
            strat_1 = experimetrics.stratselect(prob_10)
            # Select strategy of player 2
            strat_2 = experimetrics.stratselect(prob_20)

            # Determine payoffs
            if strat_1 == 0 and strat_2 == 0: # Strategy 1 / Strategy 1
                n_00 = n_00+1
                payoff_1 = S[0]
                payoff_2 = S[1]
            elif strat_1 == 0 and strat_2 == 1: # Strategy 1 / Strategy 2
                n_01 = n_01+1
                payoff_1 = S[2]
                payoff_2 = S[3]
            elif strat_1 == 1 and strat_2 == 0: # Strategy 2 / Strategy 1
                n_10 = n_10+1
                payoff_1 = S[4]
                payoff_2 = S[5]
            elif strat_1 == 1 and strat_2 == 1: # Strategy 2 / Strategy 2
                n_11 = n_11+1
                payoff_1 = S[6]-cost_1
                payoff_2 = S[7]-cost_2

            # Update payoff sum
            sum_payoff_1 = sum_payoff_1+payoff_1
            sum_payoff_2 = sum_payoff_2+payoff_2
            # Print output to .csv file and save data
            if EXTLOG == True:
                dataWriter.writerow([sim, period, attrac_10, attrac_11, attrac_20, attrac_21,
prob_10, prob_11, prob_20, prob_21, strat_1, strat_2, payoff_1, payoff_2])

            # Update attractions depending on the selected learning model
            if LM == 'rl':
                attrac_10, attrac_11, attrac_20, attrac_21 = experimetrics.rl(phi, strat_1, strat_2,
payoff_1, payoff_2, attrac_10, attrac_11, attrac_20, attrac_21)
            elif LM == 'wfpm':
                # Potential payoffs of player 1
                if strat_2 == 0:
                    payoff_10 = S[0]
                    payoff_11 = S[4]
                elif strat_2 == 1:
                    payoff_10 = S[2]
                    payoff_11 = S[6]-cost_1
                # Potential payoffs of player 2
                if strat_1 == 0:
                    payoff_20 = S[1]
                    payoff_21 = S[3]
                elif strat_1 == 1:
                    payoff_20 = S[5]
                    payoff_21 = S[7]-cost_2
                # Update attractions
                attrac_10, attrac_11, attrac_20, attrac_21, exp = experimetrics.wfpm(phi, payoff_10,
payoff_11, payoff_20, payoff_21, attrac_10, attrac_11, attrac_20, attrac_21, exp)

    if EXTLOG == True:
        dataFile.close()

    return n_periods, n_00, n_01, n_10, n_11, sum_payoff_1, sum_payoff_2, sum_cost_1, sum_cost_2,
sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21
```

A.7 Simulation Series CDG (simulations_matrix_phi-lam-series.py)

```
import math
import csv
import time
import os
import matrixgame
import sys
from tqdm import trange # Import tqdm progress bar
from config import S_MAT, N, M_MAT, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, EXTLOG

TIME = time.strftime("D%Y%m%dT%H%M%S") # Date/time of simulation
LM = str(sys.argv[1]) # Get command line argument 1
print('Learning model: ' + str(sys.argv[1])) # Print used learning model to console

# Create file for summary table
os.makedirs('output/' + TIME + '-' + LM)
summaryTableFile = open('output/' + TIME + '-' + LM + '/' + 'summaryTable-' + TIME + '.csv', 'w',
newline='') # Create summary table file
summaryTableWriter = csv.writer(summaryTableFile) # Create data file writer
summaryTableWriter.writerow(['lambda', 'phi', 'n_periods', 'n_ws', 'n_wr', 'n_is', 'n_ir',
                             'sum_payoff_1', 'sum_payoff_2', 'average payoff 1', 'average payoff
2','average cost 1', 'average cost 2',
                             'average prob_10 (w)', 'average prob_11 (i)', 'average prob_20 (s)',
'average prob_21 (r)'])

# Run simulations for lambda between 0 and 1
for lam in trange(0,11,desc='Lambda'):
    lam=lam/10 # range() function only supports integers, convert integer to respective float

    if EXTLOG == True:
        os.makedirs('output/' + TIME + '-' + LM + '/lam' +str(lam)) # Create sub-directories for each
lambda

    # Run simulations for phi between 0 and 1
    for phi in trange(0,11,desc='Phi   '):
        phi=phi/10 # range() function only supports integers, convert integer to respective float

        (n_periods, n_ws, n_wr, n_is, n_ir, sum_payoff_1, sum_payoff_2, sum_cost_1, sum_cost_2,
         sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21) = matrixgame.twobytwo(S_MAT, N, M_MAT,
C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM)

        # Write summary to summary table (1 line for each lam/phi combination)
        summaryTableWriter.writerow([lam, phi, n_periods, n_ws, n_wr, n_is, n_ir,
                                     sum_payoff_1, sum_payoff_2, sum_payoff_1/n_periods,
sum_payoff_2/n_periods, sum_cost_1/n_periods, sum_cost_2/n_periods,
                                     sum_prob_10/n_periods, sum_prob_11/n_periods,
sum_prob_20/n_periods, sum_prob_21/n_periods])

# Close summary table file and print statement of completion
summaryTableFile.close()
print('\nCompleted ' + str(N) + ' simulations for every phi and lambda between 0 and 1 iterated at
0.1')
```

A.8 Simulation CDG (simulations_matrix_phi-lam-static.py)

```
import math
import csv
import time
import os
import matrixgame
import sys
from config import S_MAT, N, M_MAT, C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, EXTLOG

TIME = time.strftime("D%Y%m%dT%H%M%S") # Date/time of simulation
LM = str(sys.argv[1]) # Get command line argument 1
print('Learning model: ' + str(sys.argv[1])) # Print used learning model to console

if EXTLOG == True:
    os.makedirs('output/' + TIME + '-' + LM + '/lam' +str(lam)) # Create sub-directory for lambda

    (n_periods, n_ws, n_wr, n_is, n_ir, sum_payoff_1, sum_payoff_2, sum_cost_1, sum_cost_2,
     sum_prob_10, sum_prob_11, sum_prob_20, sum_prob_21) = matrixgame.twobytwo(S_MAT, N, M_MAT,
C_1_MIN, C_1_MAX, C_2_MIN, C_2_MAX, lam, phi, TIME, LM)
```

Appendix B Additional Simulation Results and Selected Data Tables

B.1 Data for CDG (RL) A and C (Figure 7 and 8)

Table 7: Data for CDG (RL) A, Probability of Strategy I

ϕ \ λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.5	0.52	0.54	0.56	0.57	0.58	0.59	0.6	0.6	0.6	0.6
0.1	0.5	0.53	0.55	0.57	0.58	0.6	0.61	0.61	0.62	0.62	0.63
0.2	0.5	0.53	0.55	0.58	0.6	0.61	0.62	0.63	0.64	0.65	0.65
0.3	0.5	0.53	0.56	0.59	0.61	0.63	0.64	0.66	0.67	0.67	0.67
0.4	0.5	0.54	0.57	0.6	0.63	0.65	0.67	0.69	0.7	0.68	0.66
0.5	0.5	0.55	0.59	0.63	0.66	0.68	0.7	0.71	0.69	0.67	0.62
0.6	0.5	0.56	0.61	0.66	0.69	0.72	0.7	0.7	0.66	0.63	0.58
0.7	0.5	0.58	0.65	0.71	0.73	0.72	0.67	0.61	0.56	0.54	0.51
0.8	0.5	0.61	0.71	0.75	0.69	0.62	0.55	0.57	0.54	0.54	0.53
0.9	0.5	0.74	0.75	0.64	0.59	0.58	0.58	0.56	0.54	0.53	0.53
1.0	0.5	0.72	0.63	0.61	0.58	0.53	0.55	0.55	0.53	0.53	0.53

Table 8: Data for CDG (RL) C, Probability of Strategy I

ϕ \ λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.5	0.52	0.55	0.56	0.58	0.59	0.6	0.61	0.62	0.62	0.62
0.1	0.5	0.53	0.55	0.57	0.59	0.61	0.62	0.63	0.64	0.65	0.65
0.2	0.5	0.53	0.56	0.58	0.61	0.62	0.64	0.66	0.67	0.68	0.68
0.3	0.5	0.54	0.57	0.6	0.62	0.65	0.67	0.68	0.69	0.71	0.71
0.4	0.5	0.54	0.58	0.62	0.64	0.68	0.7	0.72	0.73	0.73	0.74
0.5	0.5	0.55	0.6	0.64	0.68	0.71	0.74	0.76	0.77	0.76	0.78
0.6	0.5	0.56	0.62	0.68	0.72	0.76	0.78	0.79	0.8	0.83	0.86
0.7	0.5	0.58	0.66	0.73	0.77	0.8	0.83	0.87	0.91	0.94	0.97
0.8	0.5	0.62	0.74	0.81	0.85	0.91	0.96	0.98	0.99	1	1
0.9	0.5	0.75	0.85	0.91	0.96	0.97	0.98	1	1	1	1
1.0	0.5	0.85	0.91	0.94	0.96	0.98	0.99	1	0.99	1	1

Table 9: Data for CDG (RL) A, Probability of Strategy R

ϕ \ λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.5	0.5	0.5	0.49	0.49	0.48	0.48	0.47	0.46	0.46	0.45
0.1	0.5	0.5	0.5	0.49	0.49	0.48	0.48	0.47	0.47	0.47	0.45
0.2	0.5	0.5	0.5	0.49	0.49	0.49	0.48	0.47	0.47	0.47	0.47
0.3	0.5	0.5	0.5	0.49	0.49	0.49	0.48	0.47	0.48	0.48	0.48
0.4	0.5	0.5	0.5	0.49	0.48	0.49	0.49	0.5	0.48	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.49	0.49	0.5	0.5	0.51	0.49	0.49
0.6	0.5	0.5	0.49	0.5	0.5	0.51	0.52	0.5	0.47	0.47	0.47
0.7	0.5	0.5	0.5	0.5	0.53	0.52	0.51	0.48	0.46	0.45	0.46
0.8	0.5	0.5	0.51	0.53	0.51	0.49	0.49	0.47	0.51	0.48	0.49
0.9	0.5	0.5	0.5	0.49	0.52	0.5	0.51	0.49	0.46	0.5	0.48
1.0	0.5	0.51	0.5	0.52	0.51	0.5	0.5	0.51	0.46	0.48	0.5

Table 10: Data for CDG (RL) C, Probability of Strategy R

ϕ \ λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.49	0.49	0.49	0.49	0.49
0.1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.51
0.2	0.5	0.5	0.5	0.5	0.5	0.51	0.5	0.51	0.51	0.53	0.54
0.3	0.5	0.5	0.5	0.51	0.51	0.51	0.51	0.53	0.55	0.56	0.59
0.4	0.5	0.5	0.51	0.51	0.52	0.52	0.54	0.56	0.58	0.63	0.66
0.5	0.5	0.5	0.51	0.52	0.53	0.55	0.57	0.62	0.65	0.71	0.72
0.6	0.5	0.51	0.51	0.53	0.56	0.6	0.66	0.71	0.75	0.8	0.84
0.7	0.5	0.51	0.53	0.57	0.65	0.72	0.78	0.85	0.9	0.94	0.97
0.8	0.5	0.52	0.58	0.7	0.81	0.9	0.95	0.98	0.99	1	1
0.9	0.5	0.59	0.78	0.88	0.92	0.97	0.98	0.99	1	1	1
1.0	0.5	0.73	0.85	0.92	0.95	0.98	0.98	0.99	0.99	1	1

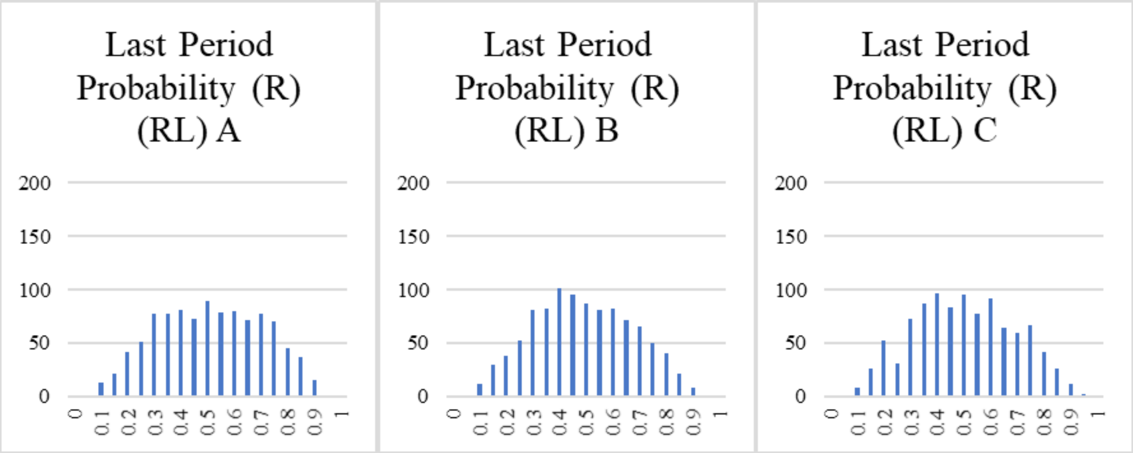
B.2 Data for CDG (BL) A (Figure 9)

Table 11: Data for CDG (BL) A, Probability of Strategy I

ϕ \ λ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	0.5	0.52	0.55	0.57	0.59	0.61	0.63	0.65	0.66	0.68	0.69
0.1	0.5	0.52	0.55	0.57	0.59	0.62	0.63	0.65	0.67	0.68	0.69
0.2	0.5	0.52	0.55	0.57	0.59	0.62	0.64	0.65	0.67	0.68	0.7
0.3	0.5	0.52	0.55	0.57	0.6	0.62	0.64	0.66	0.67	0.69	0.7
0.4	0.5	0.52	0.55	0.57	0.59	0.62	0.64	0.66	0.67	0.69	0.71
0.5	0.5	0.52	0.55	0.57	0.6	0.62	0.64	0.66	0.68	0.69	0.71
0.6	0.5	0.52	0.55	0.57	0.6	0.62	0.64	0.66	0.68	0.7	0.71
0.7	0.5	0.52	0.55	0.57	0.6	0.62	0.64	0.66	0.68	0.7	0.72
0.8	0.5	0.52	0.55	0.57	0.6	0.62	0.64	0.66	0.68	0.7	0.72
0.9	0.5	0.52	0.55	0.57	0.59	0.62	0.64	0.66	0.68	0.7	0.72
1.0	0.5	0.52	0.55	0.57	0.59	0.62	0.64	0.66	0.68	0.7	0.72

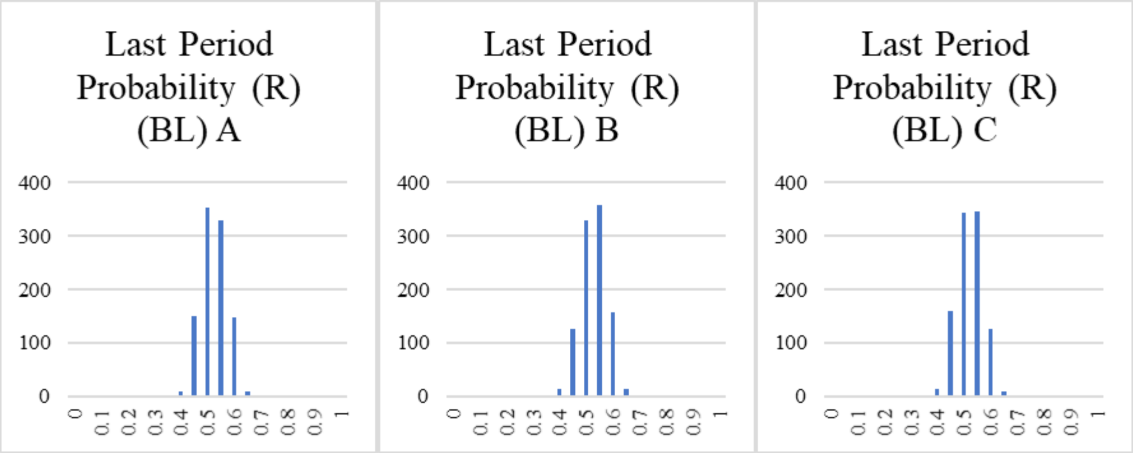
B.3 Final Probabilities of State SM Playing R in CDG (ϕ = 0.5; λ = 0.5)

Figure 16: Histograms of Final Probabilities of R in CDG (RL; ϕ = 0.5; λ = 0.5)



Own illustration.

Figure 17: Histograms of Final Probabilities of R in CDG (BL; ϕ = 0.5; λ = 0.5)

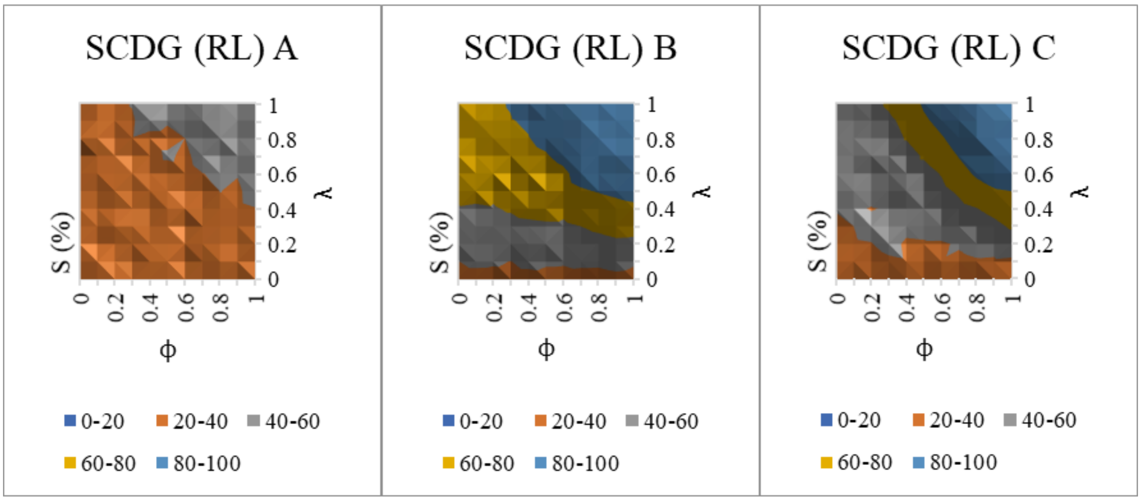


Own illustration.

Figure 14 and 15 show that P_{SM}^{R} in the last period clusters around 0.5. This effect is strong in CDG (BL) while CDG (RL) results are more randomly distributed.

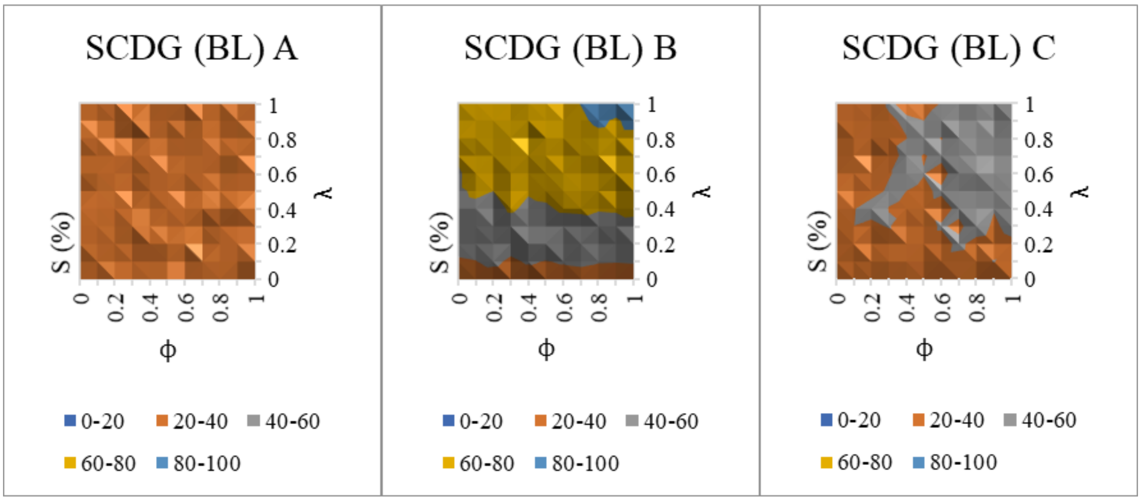
B.4 Average S (%) in SCDG Simulations

Figure 18: Influence of ϕ and λ on Average Occurrence of S in SCDG (RL)



Own illustration.

Figure 19: Influence of ϕ and λ on Average Occurrence of S in SCDG (BL)



Own illustration.

B.5 Data for SCDG (RL) A and C (Outcome W, Figure 12)

Table 12: Data for SCDG (RL) A, Average Occurrence of Outcome W

ϕ \ λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	67.5	66	66.5	67.1	65.8	65.3	64.8	67.3	60.1	64.8	66
0.1	66.3	62.4	68.5	66.3	67.2	65.6	62.5	62.9	60.2	61.2	62.2
0.2	66.6	67.5	66.7	62.9	65.5	62.2	62.6	62.4	63.6	61.8	62.4
0.3	67.2	69	66.1	63.4	65.2	65.2	62.4	64.8	60.3	60.8	59.6
0.4	66.3	64.7	65.4	65.4	65.6	63.1	64	63.3	62.7	55.8	59.6
0.5	67.3	64.1	65.3	64	62.1	62.7	64	58.7	62.8	58.9	59.5
0.6	69.1	69.8	65.4	67.6	62.8	64.9	62.4	61.5	60	57.1	56.3
0.7	68.5	66.1	64.3	63.3	63.6	64.2	60.1	57.2	57.6	53.2	53.2
0.8	67.1	63.3	63.9	62.6	61.2	59.9	59.1	60	54.5	54.2	53.6
0.9	69.3	64.8	65.3	63.2	61.1	60.9	59.7	58.1	58.5	56.1	56.1
1.0	66.9	67.5	62.3	62.9	60.2	57.5	58.4	55.1	56.3	53.5	53.2

Table 13: Data for SCDG (RL) C, Average Occurrence of Outcome W

$\phi \backslash \lambda$	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	64.9	62	63.7	63.6	59	55.2	53.3	54.2	54	54.3	52.4
0.1	68.5	62.2	61	60	58.4	54	53.3	51.5	51.7	51.6	48.4
0.2	67.6	64.4	60.4	56.8	60.5	56.9	54.4	47.3	49.4	47.5	45.1
0.3	67.7	61.8	56.7	58.5	59.1	54.2	51.3	47.3	46.6	43	37.3
0.4	65.8	60.7	61.4	57.4	56.2	49.1	48	47.6	39	35.7	28.4
0.5	67.4	63.4	60.9	57.4	54.6	51.3	47.3	37.9	30.6	22.4	17.6
0.6	66	60.4	61	55.6	54.2	44.9	38.9	28.7	21.3	15.3	11.9
0.7	68.1	62	59	53.1	45.5	37	26.2	19.1	13.9	9.4	6.7
0.8	66.8	60.8	57	53.5	40.8	26.7	17.5	13.9	6.9	4.8	2.9
0.9	64.3	61.2	56.7	46.3	32.7	21.2	14.1	8.7	6	3.5	2.4
1.0	64.2	62.2	52.9	36.9	28.9	19.9	11.4	7.9	6.1	3	3.1

B.6 Data for SCDG (BL) A and C (Outcome W, Figure 13)

Table 14: Data for SCDG (BL) A, Average Occurrence of Outcome W

$\phi \backslash \lambda$	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	68.7	69.1	65.9	65.1	64.5	68.4	66.1	67.9	64.9	65.5	67.4
0.1	69	66.4	66.7	65.9	65.7	66	64.9	66.4	65.4	68	66
0.2	66.3	68	65.8	66.4	66.3	66.7	65.8	65.3	68	67.5	66.5
0.3	67.5	67.8	67.2	63.5	65.7	69.3	68	65.7	69.4	62.7	68.7
0.4	65.7	66.7	63.6	65.1	66.9	65.9	65.9	65.8	66.3	67	67.2
0.5	64.9	65.5	62.6	66	65.3	65.8	65.3	68.2	68.2	65.4	64.6
0.6	68.4	69	65.2	65	65.4	68.3	66	66.4	66.5	67.1	62.8
0.7	67.3	65.5	69.7	69.3	64.3	67.1	66.5	65.6	64.6	64.4	66.4
0.8	67.3	66.1	66.6	67.1	65	68.1	68.8	66	67.1	66.6	65
0.9	70.7	66.5	65.6	67.1	61.7	67.1	67	64.2	66.5	63.8	66.3
1.0	69.3	64.3	67.2	67.6	65.6	66.8	66.5	68.5	65.5	66.7	63.9

Table 15: Data for SCDG (BL) B, Average Occurrence of Outcome W

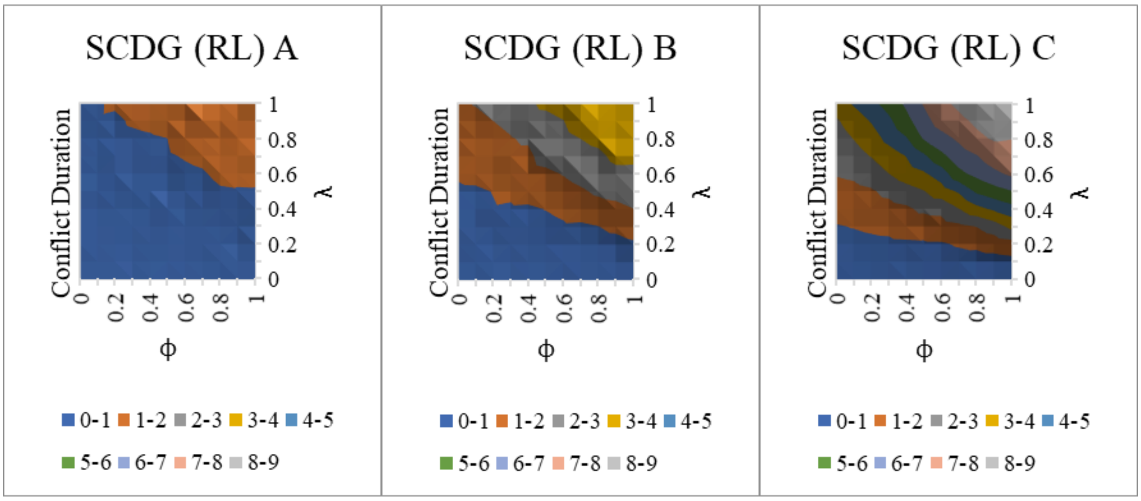
$\phi \backslash \lambda$	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	67.9	61.5	55.7	49.1	42.3	41.4	41.5	38.4	34.1	35.9	35
0.1	66	60.4	55.2	48.4	41.7	38.7	34.9	35.9	32.8	32.4	33.1
0.2	67.6	57	53.8	48.1	45.3	40	37.2	33.1	31.6	29.9	31.1
0.3	67.5	62.5	54.2	45.8	38	36.2	32.5	30.9	28.8	28.3	28.7
0.4	68.2	58.5	50.3	48.4	42.9	38.8	34.9	29.1	31.6	26.1	24.4
0.5	66	60.6	50.5	47.7	42	36.3	32.8	28.6	24.8	25.4	24.2
0.6	66.3	58.8	53.5	47.5	37.7	33.6	31.1	30.2	25.6	24	25.6
0.7	64.5	58	50.3	46.8	37	33.5	31.2	26.7	24.1	21.3	19.7
0.8	67.1	59.2	53.6	44.8	39.1	32.1	29.8	24.5	22.2	18.4	19.2
0.9	67.8	59.5	52.7	45	38.6	35.4	26.5	27.4	22.6	21	15
1.0	67.8	59	50	46	35	29.8	27.9	22.8	22.7	17	16.7

Table 16: Data for SCDG (BL) C, Average Occurrence of Outcome W

$\phi \backslash \lambda$	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	68.5	62.8	63	62.2	62.2	63.8	65.9	66.7	67.1	65	65.6
0.1	68.6	65.5	60.1	60.1	60.3	64.7	64.9	63.2	63.3	64.3	64.4
0.2	67.3	65.2	65.7	61.4	57.8	61.8	63.8	66.7	62.9	62.6	63.2
0.3	66.2	62.3	63.6	59.6	60.1	59.3	60.4	62.6	61.5	60.6	59.4
0.4	66.9	62.6	62.1	62.8	60.5	60	58.2	58.1	60.1	59.4	61.3
0.5	68.2	64.7	62.1	62.3	61.2	61	59.8	60.3	57.7	60	62.5
0.6	67.1	61.7	62.4	59.3	62.6	59	62.1	58.6	60.2	56.7	59.8
0.7	65.9	62.4	58.5	61.3	58.3	58.7	56.5	57.5	56.6	58.4	57.4
0.8	67	63	62.4	56.4	58.3	58.6	54.9	57.4	58.1	58	56
0.9	66.8	59.7	63.2	61.2	59.6	57.8	56.9	58.6	57.7	59.6	58.4
1.0	66.2	64.1	61.6	57.7	58.5	58.7	58.2	58.3	57.4	54.8	57.1

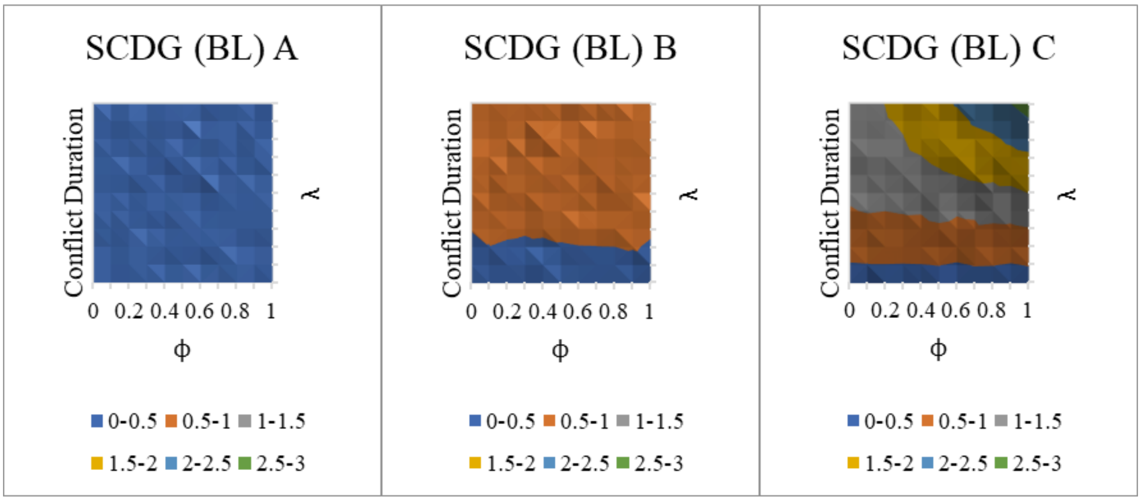
B.7 Average T (%) and Conflict Duration in SCDG Simulations

Figure 20: Influence of ϕ and λ on SCDG (RL) Conflict Duration



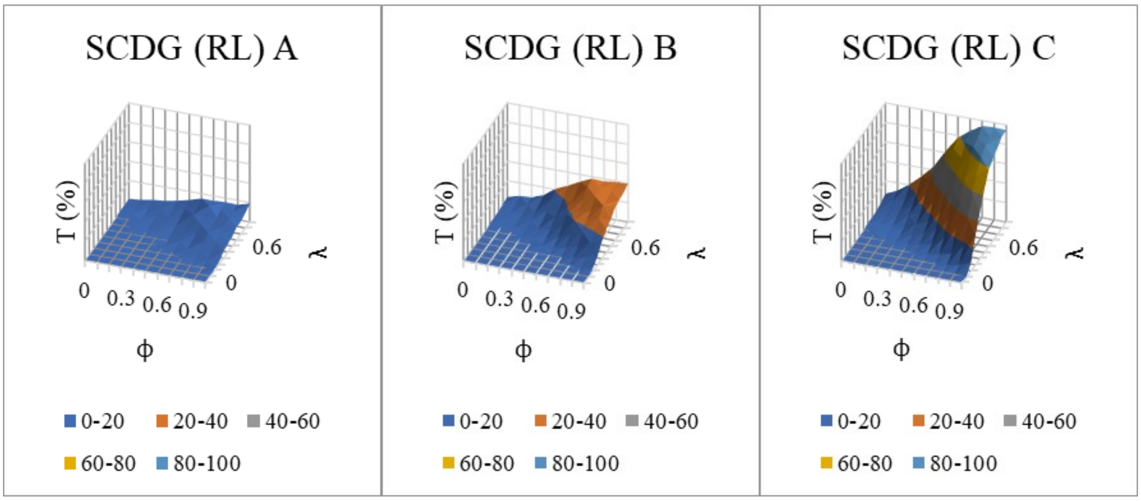
Own illustration.

Figure 21: Influence of ϕ and λ on SCDG (BL) Conflict Duration



Own illustration.

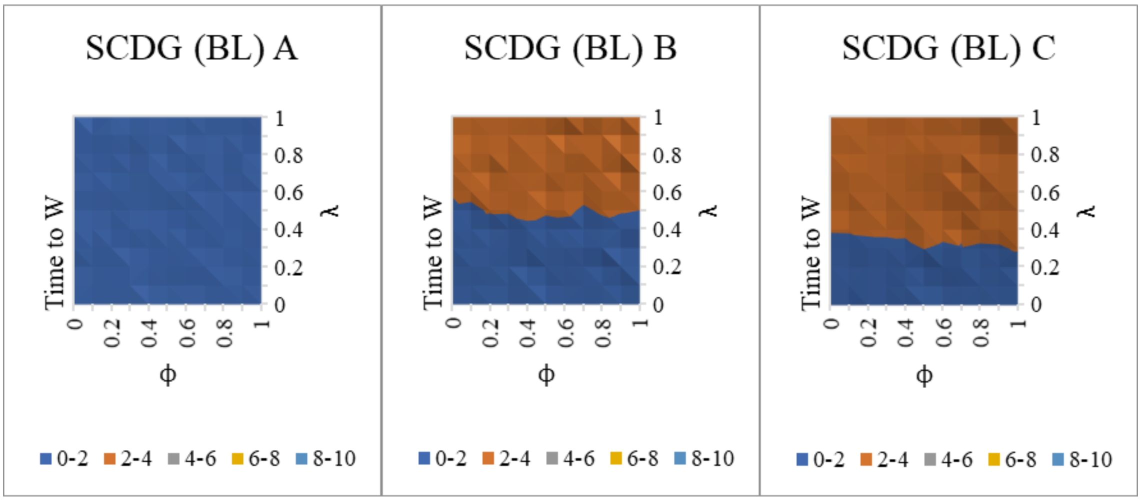
Figure 22: Influence of ϕ and λ on Average T (%) in SCDG (RL) Simulations



Own illustration.

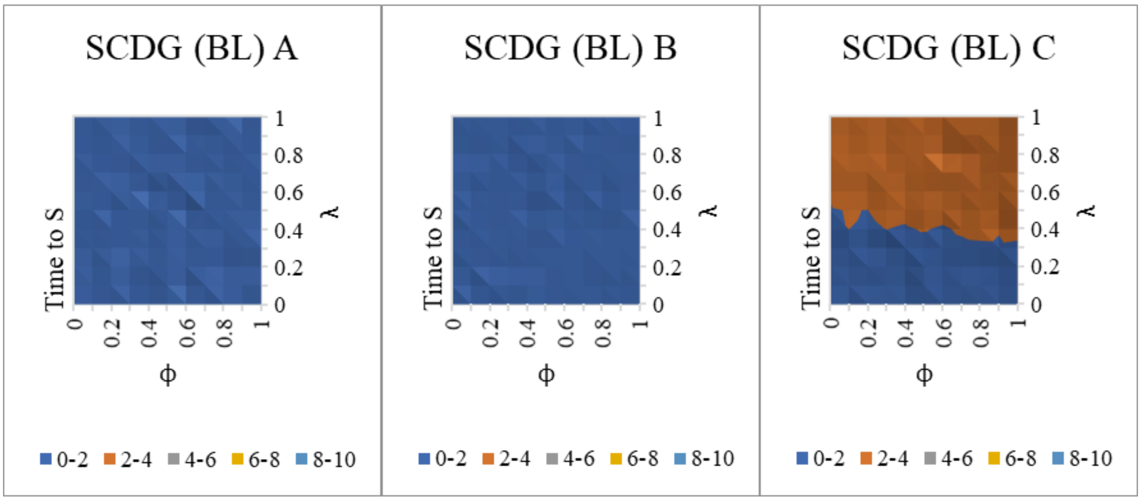
B.8 Average Time to W and S in SCDG (BL) Simulations

Figure 23: Influence of ϕ and λ on the Average Time to W in SCDG (BL)



Own illustration.

Figure 24: Influence of ϕ and λ on the Average Time to S in SCDG (BL)



Own illustration.

B.9 Data for SCDG (RL) C (Time to W and S, Figure 14 and 15)

Table 17: Data for SCDG (RL) C, Average Time to W

ϕ \ λ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.0	1.3529	1.5597	1.7598	1.9481	2.3797	2.7047	2.9737	3.2638	3.4778	3.6372	3.9981
0.1	1.3445	1.5338	1.7689	2.1133	2.4709	2.7148	3.0300	3.4563	3.7118	3.9690	4.4855
0.2	1.3151	1.5745	1.8460	2.0792	2.6000	2.9104	3.3493	3.7759	4.1235	4.4505	4.8204
0.3	1.3117	1.5485	1.8466	2.2838	2.7360	3.1421	3.8519	4.1691	4.7060	4.6791	4.8579
0.4	1.3374	1.5091	1.8990	2.4112	2.8648	3.3625	4.1750	4.2857	4.5769	4.9496	5.4366
0.5	1.3175	1.6136	1.8801	2.2439	3.0586	3.5478	4.0655	4.1557	4.6667	5.0268	5.3580
0.6	1.3182	1.5861	1.9016	2.5414	2.8875	3.5880	3.8817	4.5889	4.3756	5.0065	4.7311
0.7	1.3142	1.6419	2.0153	2.4633	2.9582	3.3649	3.5038	3.7644	3.5683	3.3617	3.5672
0.8	1.3159	1.5855	2.0877	2.4542	2.8799	2.9401	2.7371	2.5755	2.2609	2.0833	2.1034
0.9	1.3359	1.6013	1.9559	2.2808	1.9908	2.0943	1.7872	1.3793	1.5833	1.6286	1.3333
1.0	1.3193	1.6334	1.7675	1.6938	1.7301	1.6131	1.4737	1.3418	1.3770	1.3667	1.2581

Table 18: Data for SCDG (RL) C, Average Time to S

$\phi \setminus \lambda$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	1.2849	1.5500	1.6832	1.9533	2.2854	2.6339	3.1842	3.2773	3.6174	4.1707	4.5693
0.1	1.3048	1.5238	1.7205	1.9200	2.3269	2.8152	3.3490	3.7794	4.3685	4.9525	5.7248
0.2	1.2901	1.4888	1.7828	2.2014	2.6203	2.8933	3.5877	4.2941	4.8933	5.6514	6.4918
0.3	1.2910	1.5550	1.8406	2.0867	2.6284	3.3341	4.1499	4.7875	5.8652	6.6386	7.3190
0.4	1.3421	1.5496	1.8161	2.4061	2.8128	3.8507	4.8654	5.7233	6.7918	7.6936	7.8897
0.5	1.3436	1.5683	1.9412	2.5141	3.4361	4.3943	5.6034	6.8647	8.1787	8.3673	8.7051
0.6	1.3147	1.5985	1.9333	2.7297	3.7162	5.3593	6.5336	7.6662	8.5197	9.1429	9.3848
0.7	1.3919	1.6316	2.2000	3.1791	4.8624	6.1952	7.9282	8.3956	8.9477	9.3863	9.5766
0.8	1.4127	1.6122	2.3674	3.6839	5.9747	7.3602	8.2497	8.8200	9.3244	9.5011	9.7951
0.9	1.3501	1.6392	3.0600	4.6052	6.7415	7.8401	8.7695	9.2662	9.5691	9.6363	9.7951
1.0	1.3296	1.8360	3.5414	5.7765	7.2377	8.3084	9.0824	9.2812	9.5240	9.7639	9.7998

B.10 Random Strategy Selection with Absent ‘Initial Attraction’

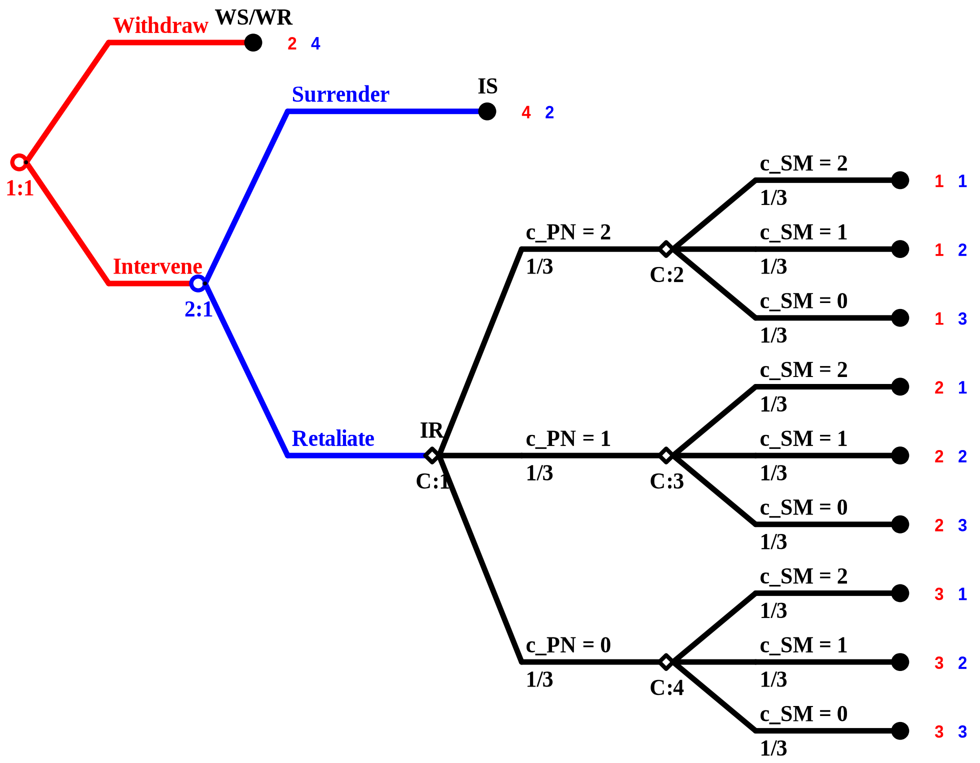
$$A_i^1(t-1)=0\,;\,A_i^0(t-1)=0$$

$$P_i^j(t)=\frac{\exp\left(\lambda A_i^j(t-1)\right)}{\exp\left(\lambda A_i^1(t-1)\right)+\exp\left(\lambda A_i^0(t-1)\right)}=\frac{\exp(\lambda 0)}{\exp(\lambda 0)+\exp(\lambda 0)}=\frac{1}{1+1}=0.5$$

Appendix C Gambit 15.1.1

C.1 Cybered Deterrence Game in Gambit 15.1.1

Figure 25: Extensive Form of the CDG with Chance Player (Gambit 15.1.1)



Own illustration, extensive form of the ‘Cybered Deterrence Game’ with chance player in Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014). The corresponding Gambit file (CyberedDeterrenceGame_ChancePlayer.gbt) can be found in the GitLab repository belonging to this paper (Gerritzen, 2019).

C.2 Data of the Quantal Reponse Equilibrium of the CDG

Table 19: Data of the Quantal Response Equilibrium of the CDG (Gambit 15.1.1)

λ_{QRE}	W	I	R	S
0	0.5	0.5	0.5	0.5
0.024494	0.49	0.51	0.5	0.5
0.051436	0.49	0.51	0.5	0.5
0.081069	0.48	0.52	0.5	0.5
0.113658	0.47	0.53	0.5	0.5
0.149494	0.46	0.54	0.5	0.5
0.188894	0.45	0.55	0.5	0.5
0.232207	0.44	0.56	0.5	0.5
0.279809	0.43	0.57	0.5	0.5
0.332111	0.42	0.58	0.5	0.5
0.389559	0.40	0.60	0.5	0.5
0.452635	0.39	0.61	0.5	0.5
0.521857	0.37	0.63	0.5	0.5
0.597785	0.35	0.65	0.5	0.5
0.681018	0.34	0.66	0.5	0.5
0.772198	0.32	0.68	0.5	0.5
0.872008	0.29	0.71	0.5	0.5
0.981182	0.27	0.73	0.5	0.5
1.100503	0.25	0.75	0.5	0.5
1.230815	0.23	0.77	0.5	0.5
1.37303	0.20	0.80	0.5	0.5
1.528145	0.18	0.82	0.5	0.5
1.697256	0.15	0.85	0.5	0.5
1.881588	0.13	0.87	0.5	0.5
2.082511	0.11	0.89	0.5	0.5
2.301573	0.09	0.91	0.5	0.5
2.540529	0.07	0.93	0.5	0.5
2.801365	0.06	0.94	0.5	0.5
3.086327	0.04	0.96	0.5	0.5
3.397941	0.03	0.97	0.5	0.5
3.73904	0.02	0.98	0.5	0.5
4.112772	0.02	0.98	0.5	0.5
4.522629	0.01	0.99	0.5	0.5
4.972455	0.01	0.99	0.5	0.5
5.466471	0	1	0.5	0.5
6.009299	0	1	0.5	0.5
6.605991	0	1	0.5	0.5
7.262073	0	1	0.5	0.5
7.983585	0	1	0.5	0.5
8.777145	0	1	0.5	0.5
9.650002	0	1	0.5	0.5
10.610116	0	1	0.5	0.5
11.666227	0	1	0.5	0.5
12.827944	0	1	0.5	0.5
14.105829	0	1	0.5	0.5
15.511503	0	1	0.5	0.5

Source: Gambit 15.1.1 (McKelvey, McLennan and Turocy, 2014). The corresponding Gambit file and the full QRE calculation output of Gambit (CyberedDeterrenceGame_ChancePlayer_Qre.csv) can be found in the GitLab repository belonging to this paper (Gerritzen, 2019).

Previous Discussion Papers

THE END OF INTERVENTIONS? – Simulating Cyberweapons as Deterrence Against Humanitarian Interventions
No. 6
Christopher Gerritzen
Januar, 2020
PDF, 85 Seiten

CROWDWORKING MONITOR NR. 2 – für das Verbundprojekt "Crowdworking Monitor"
No. 5
Prof. Dr. Oliver Serfling
Februar, 2018
PDF, 46 Seiten

CROWDWORKING MONITOR NR. 1 – für das Verbundprojekt "Crowdworking Monitor"
No. 4
Prof. Dr. Oliver Serfling
September, 2018
PDF, 50 Seiten

ZUR MESSUNG VON FRAKTIONSMACHT – Ein gewichteter Machtindex mit Kohäsionskoeffizienten zur Messung der relativen Fraktionsmacht in Parlamenten am Beispiel des Deutschen Bundestages Empirical Evidence for Developing and Developed Countries
No. 3
Jakob Lempp and Thomas Pitz
Januar, 2017
PDF, 16 Seiten

THE ROLE OF GOVERNANCE ON PROMOTING LONGER, HEALTHIER LIVES – Empirical Evidence for Developing and Developed Countries No. 2
Oliver Serfling and Zunera Rana
August, 2015
PDF, 38 Seiten

EYE-TRACKING IN BEHAVIOURAL ECONOMICS AND FINANCE – A Literature Review
No. 1
Jörn Sickmann and Ngan Le
September, 2016
PDF, 40 Seiten