CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Political Studies

Department of Security Studies

Master’s Thesis

2024 Adam Voldřich

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Political Studies

Department of Security Studies

Adam Voldřich

Automating Compliance:

The Role of Machine Learning in International Prohibition Regimes

Master Thesis

Author of the Thesis: Adam Voldřich

Study Programme: Security Studies

Supervisor: Mgr. Petr Špalda, Ph.D.

Year of defence: 2024

Bibliographic Note

Voldrich, Adam Automating Compliance: The Role of Machine Learning in International Prohibition Regimes, Prague 2024. Master's thesis (Mgr). Charles University, Faculty of Social Sciences, Institute of Political Studies, Department of Security Studies. Supervisor Mgr. Petr Špelda, Ph.D.

Length of the Thesis: 99 960 (inc. spaces)

Abstract

This thesis explores the potential of machine learning in enhancing the enforcement of cluster and nuclear international regimes. Machine learning might prove to be a powerful tool for processing large amounts of information and could enhance the capabilities of non-state actors to increasingly participate in controlling regime compliance.

The theoretical foundation of this thesis is built upon three key pillars: regime theory, machine learning and constructivism. Regime theory provides a framework for understanding the core mechanism that create both anti-proliferation and prohibition regimes. The aim of this thesis is to combine regime theory with machine learning to explore the ramifications of potential deployment of machine learning within those regimes.

The results from the analysis suggest that machine learning could significantly help with monitoring prohibition regimes. Its potential lies especially in the field of open-source intelligence, where it enables to process the cast amounts of data. This could in turn significantly help non-state entities like NGOs and their campaigns

Therefore, in this thesis it is argued that the utilisation of machine learning especially by non-governmental organisation enables them to increasingly influence prohibition regimes by shifting narratives by effectively exposing perpetrator’s defection from the regime and instating proportional retaliation in time.

Abstrakt

Tato diplomová práce zkoumá potenciál strojového učení při posilování prosazování mezinárodních režimů klastrové a jaderné munice. Strojové učení by mohlo být mocným nástrojem při zpracovávání velkého množství dat. Čímž by mohlo zvýšit kapacity a možnosti nestátních aktéru se více podílet na kontrole dodržování režimů.

Teoretický základ práce je založen na třech klíčových pilířích: teorie režimů, strojového učení a konstruktivismu. Teorie režimů tvoří rámec pro pochopení klíčových mechanismů, které tvoří jak režimy zákazu, tak režimy proti proliferaci. Cílem práce je zkombinovat teorii režimů se strojovým učením a zkoumat dopady potenciálního nasazení strojového učení v těchto režimech.

Výsledky analýzy naznačují, že strojové učení by mohlo významně pomoci při monitorování režimů. Jeho potenciál spočívá zejména v oblasti open source zpravodajství, umožnění zpracování velkých množství dat. To by mohlo výrazně pomoci nestátní subjektům, jakou jsou nevládní organizace.

Využití strojového učení a to zejména nevládními organizacemi, umožňuje zvyšovat vliv na ustanovené režimy, účinným odhalováním porušení režimů a včasným zavedením přiměřené odvety.

Key Words

Machine learning, cluster munitions, nuclear weapons, regime theory, constructivism, humanitarian disarmament, prohibition regime, non-proliferation regime

Klíčová slova

Strojové učení, kazetová munice, jaderné zbraně, teorie režimů, konstruktivismus, humanitární odzbrojení, prohibiční režim, anti-proliferační režim

Declaration of Authorship

1. The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

2. The author hereby declares that all the sources and literature used have been properly cited.

3. The author hereby declares that the thesis has not been used to obtain a different or the same degree.

In Prague on 30th 2024 Adam Voldřich

Acknowledgement

I would like to express my gratitude to my supervisor Mgr. Petr Špalda, Ph.D. for his good will and flexibility.

Obsah

Abbreviations

ABM – Anti-Ballistic Missile Treaty

ANN – Artificial Neural Network

CCM – Convention on Cluster Munitions

CMC – Cluster Munition Coalition

CM – Cluster Munitions

CNN – Convolutional Neural Networks

FFN – Feedforward Neural Network

GAN - Generative Adversarial Networks

ML – Machine Learning

NGO – Non-governmental Organization

NPT – Non-Proliferation Treaty

RNN – Recurrent Neural Networks

SALT – Strategic Arms Limitation Talks

START – Strategic Arms Reduction Treaty

TPNW – Prohibition of Nuclear Weapons

UN – United Nations

Introduction

An era characterized by rapid technological advancements and complex geopolitical challenges creates opportunities for the integration of machine learning automation in various new environments. Nuclear and cluster munitions regimes remain a critical part of global security. Both non-proliferation and prohibition regimes, aim at mitigating the devastating effects of such weapons. During the last decades, there has been a significant increase in the amount of information created and publicly available regarding cluster and nuclear weapons. This increase presents challenges due to limited capacities for data analysis and opportunities for new emerging technologies that could help manage the quantity of data available.

This thesis explores the potential of machine learning in enhancing the enforcement of these international regimes. Machine learning might prove to be a powerful tool for processing large amounts of information and could enhance the capabilities of non-state actors to increasingly participate in controlling regime compliance.

The theoretical foundation of this thesis is built upon three key pillars: regime theory, machine learning and constructivism. Regime theory provides a framework for understanding the core mechanism that create both anti-proliferation and prohibition regimes. The aim of this thesis is to combine regime theory with machine learning to explore the ramifications of potential deployment of machine learning within those regimes. Furthermore, the role of international cooperation and the influence of global norms, as emphasized by constructivism, will be examined, so that one could understand how they can support or hinder the integration of ML and if ML could empower actors to increase their influence on the regime. Technological limitation associated with ML deployment of machine learning will be discussed as well.

This thesis aims to contribution to the ongoing discourse on how emerging technologies could be harnessed in international security studies. The central argument is built on two case studies of cluster munitions and nuclear weapons and their intersection with machine learning.

Methodology

The following methodology was used to answer the research question:

How could machine learning help with enforcing prohibition and non-proliferation regimes?

This research aims to provide an overview of opportunities for machine learning and how ML could be leveraged to control compliance in prohibition regimes and non-proliferation regimes.

Constructivist ontological position and interpretivist epistemology allow one to focus on the role of social actors. A qualitative approach has been chosen as a methodological approach since it allows examination of both the benefits that ML implementation would bring to actors while also focusing on the theory of ML and its ramifications for the study. Furthermore, the qualitative approach provides a more comprehensive look at both ML implementation and the role of regime theory.

The case study method is particularly suitable as it allows for rich contextual analysis, specifically enabling the combination of perspectives on security, technology and regime theory(Gerring 2017; Maxwell 2013). This facilitates a combination of knowledge in the fields of machine learning, (i.e. the technological challenges and opportunities) and prohibition regimes in international systems. By combining knowledge from these fields, the goal is to develop theoretical framework supported by practical examples. Apart from answering the research question this thesis aims to offer a functional overview for leveraging machine learning in the enforcement of international regimes.

The grand theory that will encompass much of the discussion is Regime Theory which will allow one to explore the underlying theories of international prohibition regimes. Regime Theory is instrumental in understanding the establishment, maintenance, and enforcement of international norms and rules, particularly in the context of humanitarian disarmament regimes like cluster munitions, anti-personnel landmines, nuclear proliferation, and general arms control. (Hynek 2018; Jervis 1982)

Data Collection and Research Structure

Firstly, a literature overview was conducted to explore regime theory and identify prohibition regimes that could be suitable for analysis. The case study of cluster munitions and nuclear weapons were chosen for several reasons. First, each regime is functionally different: the regime on CM was created relatively recently, based on humanitarian principles and from the onset strongly influenced by NGOs and the public. Furthermore, cluster munitions are still actively used

in conflicts today. The nuclear weapons non-proliferation regimes on the other hand, originated mainly from state needs and priorities. Nuclear weapons have not been used in conflict since World War 2 and there is a strong taboo against the use of nuclear weapons.

After completing the literature overview of regime theory, focus shifted to literature on machine learning. This involved recognizing its core features and weaknesses and researching on how and if machine learning is used to monitor individual parts of each prohibition regime[1]. Following the literature review and examination of both cases an attempt was made to create a nexus between ML and regime theory in order to answer the research question.

Limitations

Given the scope of this research many issues and areas had to be omitted. This was done for several reasons, first author's attempt to balance ML and regime theory so they complement each other and for purely practical reasons. The choice of which issues to include in the case study was made arbitrarily by the author, although a great deal of though was put into deciding what to include and what to omit. In order to connect ML and regime theory, a balance needed to be struck between analytical depth and the breath of the covered issue area.

[1] See the division of research findings

Literature Review

Regime Theory

Regime theory is a fundamental framework in international relations that illuminates the reasons behind international state cooperation and looks into how and why states cooperate in addressing global issues. This chapter will focus on theoretical foundations of regime theory and its development, particularly in the field of international relations. By aiming at the examination of various theoretical perspectives, one should understand how international norms and rules are established, maintained, and enforced.

Initially, this part will discuss the classification of regime theory development throughout time, in the discipline of international relations. The goal of this is to illuminate the two main issues when considering international regimes. First issue is the assumptions about the nature of international relations which clarify the relationship among the fundamental key concepts in this theory, i.e. power, interests, norms. *"Second, what is the relationship between regimes and related outcomes and behaviour?"* (Krasner 1983, 1)

Hynek (2017) categorises "*three waves of scholarship*". The first, "consequentialist regime theories", that had their peak between 1970s and 1980s, is characterized as a *"theoretical convergence between neoliberal institutionalism and neorealism"* (2017, 13). The second wave: "cognitivism and theories of regimes", which centres around constructivism, cognitivism, distincts itself from the previous generation and embraces multilateralism and bureaucratic expertise. The third wave of "radical constructivist/post-structuralist", which is especially unique in its distinction from the other two waves, is characterised by Foucauldian understanding of theory. (Hynek 2017)

First Wave

The year 1970 represented a shift from state-centric realist thinking to more holistic approach connected to the liberal theories and focus on transnational actors; low politics became more salient as the world was becoming more and more interconnected, slowly leading towards prevailing liberalism (Hasenclever, Mayer, and Rittberger 2000).

The first wave, which could be related with Krasner's understanding of regimes. For Krasner (1983) there are four main characteristics. The first two: principles and norms that create the fundamental building blocks of a regime, those qualities should make them

stable. The second two principles: rules and decision-making processes that function within the set boarders determined by norms and principles, yet fluid enough to change within the system. Therefore, indicating that although there were once set practices, they may change but the rules of the "game" stay unaltered and the regime stays the same. Krasner argues that since principles and norms are fundamental in policy issues they are the core determinants rather that rules and procedures. *"Regimes can be defined as sets of implicit or explicit principles, norms, rules, and decision-making procedures around which actors' expectations converge in a given area of international relations"* (Krasner 1982, 186).

Robert O. Keohane (1982) joins Krasner in his analysis of international regimes in using rational choice theory for creating a model for explaining trends and tendencies. Importantly, Keohane recognises that due to different power diffusion among states and international entities, not all international regimes might be created voluntarily, arguing that powerful states could impose such regimes upon other states. Hence, in the Keohane's view, it is futile to examine and consider participants in international regimes as equals. Although these "imposed regimes" are still agreed upon, it is done so in constrains imposed by the dominant player. Keohane invokes prevalent realistic assumptions: *"world politics lacks authoritative governmental institutions, and are characterized by pervasive uncertainty"* (Keohane 1982, 332). Furthermore, there are other salient points made, such as that transaction costs are reduced by creating regimes (as oppose to ad hoc agreements) and enforceable legal liabilities are not fully established by international regimes, as reducing transaction costs and the exchange of information is more salient (Levy, Young, and Zurn 1995; Keohane 1982; Keohane and Nye 2012; Jervis 1982). Especially the point of exchanging information is particularly notable for this thesis, as, according to the theory, it reduces risk and uncertainty.

International policy coordination and the development of international regimes depend not merely on interests and power, or on the negotiating skills of diplomats, but also on expectations and information, which themselves are in part functions of the political structures of governments and their openness to one another. (Keohane 1982, 347)

Similarly to Krasner for Keohane and Nye the salient features for international regimes are "*networks of rules, norms, and procedures that regularize behaviour and control its effects"* (2012, 16). Furthermore, Keohane and Nye (2012) establish three main

characteristics of interdependence. First there are multiple formal and informal ties between governments. Second, there is absence of hierarchy among intestate issues. Third, military force is not used.

The first wave represented the "theoretical convergence" of neoliberalism and neorealism (Hynek 2018; 2017). The fact that the participants in international regimes represent rational players and that states should be utility maximisers. The driving principle behind both theories is that regimes are intentionally created in order to promote cooperation and limit states' self-help behaviour.

They diverge on other issues as to why one should cooperate and whether gains should be measured absolutely or relatively to others in the system (Hasenclever, Mayer, and Rittberger 2000).

Second Wave

Cognitivism is a second wave of regime theory that became prominent in the 1990s which marked the turn from state-centric understanding to neo-functional and neo-institutional one (Hynek 2017). Cognitivism criticised neoliberal approach as well as neorealism did, though from a different angle. Rather than criticizing the realist assumptions as they are, cognitivism targets the realist assumptions that made its way into neoliberalism. Therefore, it sets itself apart from both previous theories on the basis that actors' preferences are assumed to be given and are not properly examined, i.e. preferences are determined by the presumed rational behaviour. This in the eyes of cognitivists leads to the distortion of reality (Hasenclever, Mayer, and Rittberger 2000).

The key concept of cognitivist theory about regimes are epistemic communities which by controlling knowledge and information influences international policy decision-making processes by which new behavioural patterns are made, rendering the whole process harder to predict (Haas 1992).

There are two main strains of thought of regime analysis in cognitivism: weak and strong. The weak cognitivism tries to explain the actual behaviour of individuals and their influence on the regime formation, particularly the role of knowledge that is essential in making decisions in international regimes and how this knowledge is distributed and its distribution effects on the decision making process (Hasenclever, Mayer, and Rittberger 2000). It might be also called reflective approach as Keohane (1988) argues that scholars

of rational choice theory must eventually run into problem of social integration. Furthermore, international regimes are not often created out of nothing, there is a continuation of previous work as well as institutional context which is hardly explainable by rational maximizing utility individuals. Weak cognitivism concerns itself with actors' casual beliefs, strong cognitivism research on the other hand centres around knowledge as a central variable. They deny the notion of state being a rational actor (Hasenclever, Mayer, and Rittberger 2000; Haas 1992).

Hass (1992, 3) defines epistemic communities as a *"network of professional with recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain or issue-area".* It is important to note that epistemic communities differ from the bureaucratic bodies in regimes. Cognitivists consider that all interaction in the particular regime, no matter the previous possible self-interests of the actors, cause participants to change and alter their views on others, particularly in a positive sense. Even previously alienated norms are being internalised throughout time spend in the institutional settings (Wendt 1994; Hasenclever, Mayer, and Rittberger 2000; Buzan and Hansen 2009; Haas 1992).

Within this framework, strong cognitivism acknowledges the role-player model where state considers and acts upon the notion of what is expected in the particular situation, i.e. what is considered appropriate in the given situation. This signifies that some actors might have different roles (role dynamics) than others in a system (Hasenclever, Mayer, and Rittberger 2000). Furthermore, Hynek (2017) recognises four main cooperation areas: *"The power of legitimacy studying social fabrics of international political life and its forms and rules; the power of arguments inspired by Habermas communicative rationality and ethics; the power of identity where Self/Other binary gets at the forefront; and the power of history"* (Hynek 2017, 18).

Europe played a significant role in forming particular wave of regime theory that was the Tübingen School, from the University of Tübingen, a Tübingen Peace Research Group led by Professor Volker Rittberger. This considerably wide research focused on, apart from the pure theory, data-driven research as well as the utilization of statistical analysis, particularly focusing on East-West regimes (Hynek 2017; Underdal 1995). There are several distinctive features. The research is based on peace research which is linked to the group's focus primarily between Eastern and Western block. They saw the regimes as a

significant indicator in reducing the risk of conflict among parties in the given regime. Such regimes are one of the key pillars in creating security communities, and supported by two effects: domestic democratization and international civilization (Underdal 1995; Hasenclever, Mayer, and Rittberger 2000).

Tübingen School also created two distinct structures which distinguished agendas of political regimes; those benign issue areas and malign, while utilising game theory to further conceptualise the problem. They identified three categories of games for the problems: coordinated game, dilemma game, and Rambo game. This categorization is used to help explain and predict regime creation. While coordinated game could be categorised as a benign issue area, Rambo game would be considered a malign one (Underdal 1995).

The biggest leap in the second wave compared to the first one was the move from conceptual groundwork to more data-driven research, to bigger empirical studies. And finally the establishment of three major determinant factors for regimes: rules of the game, actor preferences, and the distribution of power (Underdal 1995).

Third Wave

The third wave is Radical Constructivist/Post-Structuralist, *"the incorporation of radical critical social and political theory to regime analysis"* (Hynek 2017, 18). This represents the latest change in the debate. The core difference is that unlike the previous two waves, the third one is not as interested in advancing the previous theorisation but rather in theorisation itself. The central concept of power and truth where big civilization discourses are being dismantled and the theory is looking into post-structuralists causes, e.g. how culture and symbolism effects state behaviour and decision-making processes. There are two instances in the case of chemical weapons regime[2] and the nuclear weapons regime[3]. In the case of Tannenwald (2012), the argument is that, indeed, deterrence is one of the factors why nuclear weapons have not been used since 1945. Factors as taboos and stigmatization play a significant role in dissuading the US from utilizing nuclear weapons.

[2]See: Price, R. M. (1995) A Genealogy of the Chemical Weapons Taboo. International Organization 49(1): 73-103. Price, R. M. (1997) The Chemical Weapons Taboo. Ithaca, ny: Cornell University Press.

[3] See: Tannenwald, N. (1999) The Nuclear Taboo: The United States and the Nor- mative Basis of Nuclear Non-Use. International Organization 53(3): 433-468

Tannenwald, N. (2007) The Nuclear Taboo: The United States and the Non-Use of Nuclear Weapons Since 1945. Cambridge: Cambridge University Press.

She considers several cases (the Korean War, the Vietnam War, and the Gulf War of 1991). The third wave represents a distinct change as it introduces a moral imperative as the baseline for its arguments. Both Price (1995) and Tannenwald (2007) agree that previous regime theories overlooked the moral baseline and the logic of taboos and stigmatisation which are inherently illogical in the eyes of utility maximizing rational player.

Prohibition Regimes

This chapter will discuss prohibition regimes as a distinct category in regime theory. Firstly, a general theory on prohibition regimes will be discussed, followed by two particular reviews on nuclear and cluster munition regimes.

As the framework for understanding regimes was discussed in previous chapter, prohibition regimes are a specific type of a regime within the given framework — such regimes are aiming at restricting or banning certain activities, goods or practices agreed upon by international community. These regimes typically involve bi/multilateral agreements and cooperation among actors in order to control and enforce the regimes.

Willingness to be part of and obey by the regimes might be often attributed to various factors from fear of punishment and adherence to one's norms and principles, one's benefit or just out of habit. Furthermore, the priority or agenda of regimes often reflects the power distribution of its members and their influence. Yet, particularly in the case of prohibition *regimes, "moral and emotional factors related to neither political nor economic advantages but instead invoking religious beliefs, humanitarian sentiments faith in universalism, compassion, conscience, paternalism, gear, prejudice, and the compulsion to proselytize can and do play important roles in the creation and the evolution of international regimes"* (Nadelmann 1990, 480).

Therefore, Nadelmann (1990) concludes that prohibition regimes are particularly complex and might reflect preferences from not only economic, political, and security interests but also previously mentioned moral interests which reflect differently as societies diverge on values and norms that cocreate these regimes. This in fact points at a crucial fact that there are both external and internal forces within societies that shape governments' decisions in creating and adhering to prohibition regimes. Nadelmann furthermore recognises several reasons why prohibition regimes appear: *"protecting interests, to deter, suppress, and punish undesirable activities, to provide for order,*

security, and justice among members of a community and to give force and symbolic representation to the moral values, beliefs, and prejudices of those who make the laws" (Nadelmann 1990, 480,481).

Prohibition regimes indeed do not require mutual benefit of all participants to be instated. Per the second wave of regime theory, we see that non-state actors, internal pressures within states, and moral imperatives play a crucial role in regime formation (Keohane 2005; Getz 2006).

Nadelmann (1990) suggest four stages to prohibition regime creation: During the initial stage the activity that is to be considered for prohibition is regarded as legitimate by the international community. This initial stage is a state domain, and primary concerns are political or strategic in nature. The second stage, the issue narrative is being redefined on moral/religious/humanitarian grounds. The result of the shift in narratives is that the issue area is deemed illegitimate, and states are put under pressure by various groups to abandon the activity. The third stage is crucial as the proponents (now including states) argue in favour of outlawing the activity. The active participation in outlawing the activity may take many forms (e.g. diplomatic pressure, economic activities or military interventions). Given that the third stage proved to be successful, the activity becomes a subject to international and criminal law and could be enforced, hence the legal recognition of international prohibition regime[4] in international and national laws which constitutes the fourth stage (Nadelmann 1990; Holmes and Winner 2007).

Cluster Munitions Prohibition Regime

Cluster munitions are explosive weapons that release smaller submunitions which lead to widespread and indiscriminate damage. These weapons proved to be a significant risk to human life, especially to civilians, due to their tendency to leave unexploded ordnance. The cluster munition prohibition regime is a framework of international efforts aimed at banning the use and production of such weapons. This overview examines the context, key provisions, and some of the impacts that this prohibition regime has, centring around the Convention on Cluster Munitions.

[4] For more about regime effectiveness see Getz, Kathleen A. 2006. 'The Effectiveness of Global Prohibition Regimes: Corruption and the Antibribery Convention'. Business & Society 45 (3): 254–281. https://doi.org/10.1177/0007650306286738.

Cluster munitions are defined as a conventional munition that is designed to disperse an explosive submunitions. The submunitions are referred to as bomblets which are capable to cover wide areas. They are designed either to be antipersonnel or anti-armour, though modern munitions could have several effects (Human Rights Watch 2010a; Docherty 2007).

Cluster monitions prior to 2006 had been used in several wars (e.g. Vietnam War and by Israelis in Lebanon) for their ability to cover vast areas with explosive force proving very effective against dispersed infantry targets. However, one can observe that the impact on humans by cluster munitions (both immediate and long-term) sparked humanitarian concerns. Prior to 2006, Rappert and Moyes (2009) recognise tree categories of states in their position towards cluster munitions. First, states that do not object to the use of cluster munitions. Second, states that support limited reforms, especially to international Humanitarian Law. Third, states that support them for strict regulation and restriction. The use of cluster munitions was often weight in the connection to the expected military utility and acceptable costs.

After 2006, one could observe a shift in attitudes towards cluster munitions. There are several reasons for that: Israeli actions in Lebanon, Afghanistan, and Iraq, all wars in which cluster munitions were used and their effect was well documented which sparked intensive campaign for the ban on cluster munitions. In 2003, the Cluster Munition Coalition was formed by NGOs. The problem with cluster munitions is twofold. First, the immediate indiscriminatory destruction it causes in wide areas, especially problematic in urban areas; currently we can see such effects in Ukraine where both sides have used cluster munitions, even though specifically problematic is Russian use in urban areas against civilian population (Human Rights Watch 2023). And second, the problem of undetonated ordnance which could cause harm to civilians long after its initial use. This led to the change in argumentation regarding the prohibition regime of cluster munition, where language of unacceptable harm to civilians was adopted and advocation for a total ban of cluster munitions was further advocated (Human Rights Watch 2010a; Lacey 2009; Rappert and Moyes 2009; Docherty 2007).

The turning point for the prohibition regime was in 2007 in Norway. With the Oslo Declaration, Norway started a process by which it committed itself to leading the efforts for the prohibition of use, transfer, and stockpiling of cluster munitions. Slowly, other

nations joined Norway in prohibition and the regime was slowly being created, building on the success of the Mine Ban Treaty which build on humanitarian disarmament, addressing arms control from a civilian perspective. The Oslo process culminated in the adoption of the Convention on Cluster Munitions 2008 in the final text in Dublin. The wide support for the ban and the humanitarian campaign, the culmination of pressures from both states, NGOs and civil society, led to a broader prohibition regime. The most discussed issues were about the definitions of cluster munitions (Human Rights Watch 2010b; Lacey 2009; Rappert and Moyes 2009). The previous definitions about banning cluster munitions based on their unacceptable harm to civilian population was abandoned in favour of a clearer definition:

"Cluster Munitions means a conventional munition that is designed to disperse or release explosive submunitions each weighing less than 20 kilograms, and includes those explosive submunitions." (United Nations 2008, 4)

There are several key provisions that CCM (Convention on Cluster Munitions) has. First, article 1 states that each signatory is obliged to never "*a) use cluster munitions b) develop, produce, otherwise acquire, stockpile, retain or transfer to anyone, directly or indirectly, cluster munitions c) assist, encourage or induce anyone to engage in any activity prohibited to a State Party under this Convention*" (United Nations 2008, 3).

There are several provisions that narrow down what is not meant as a cluster munition and hence not regulated by this treaty. Furthermore, in article 3 the treaty established obligations to destroy stockpiles of cluster munitions. Although the humanitarian language was removed from the initial definition of cluster munition, it is present throughout the document, article 4 represent a positive humanitarian obligation in requiring states to destroy any remnants of cluster munition, as well as in education and victim assistance (Human Rights Watch 2010b; United Nations 2008).

Although there are 108 signatories of the CCM, some of the major powers do not take part (the United States, China, Russia, India, Pakistan, Ukraine) which holds back the efforts of this prohibition regime. Those states "*have pursued international and national measures in the name of balancing military needs and humanitarian concerns*" (Human Rights Watch 2010b, 160). Given the current situation (in the European context, that both Russia and Ukraine use cluster munitions, and the fact that despite international pressures, the United States still possess cluster munitions), one could argue that the stigmatisation

of those weapons still have not peeked. Yet, the success of CCM and its ideological predecessor the Anti-personnel Mine Ban treaty might prove to be an important framework for creating prohibition regimes.

The Non-proliferation Regime

The non-proliferation regime refers to a series of international treaties and institutions that are designed to prevent the spread of nuclear weapons, ensure peaceful use of nuclear energy, and promote nuclear disarmament. The corner stone of this regime is the Nuclear Non-Proliferation Treaty (NPT) in 1968 from which the non-proliferation regime emerged. The regime addressed nuclear testing and proliferation, the security of nuclear material, nuclear terrorism and commercial applications. (Siracusa and Warren 2018; Egeland et al. 2018; Meyer 2017)

United Nations emerged as a central organisation in the new regime and continues to play a central role to this day. There are seven pillars that support the non-proliferation regime: 1) Atoms for Peace programme; 2) the International Atomic Energy Agency; 3) Nuclear free zones (NWFZs) 1959; 4) Limited Test Ban Treaty 1963; 5) the informal London Nuclear Supplier Group; 6) Cooperative Nuclear Threat Reduction programme 1992; 7) Missile Technology Control Regime 1987 (Siracusa and Warren 2018, 4–5)

The difference from the prohibition regime on cluster munitions is that the non-proliferation regime was created by states (the United Stated and the Soviet Union) which was the result of structural powers, the distribution of power. Although shared norms provided the regime with legitimacy it was the Soviet-US cooperation that enabled this regime. (Ruzicka 2018)

Throughout time NGOs and other non-state entities began to be more vocal and influential but it was not until TPNW treaty that changed the status quo of non-proliferation discussion about nuclear weapons, their legitimacy and claimed necessity. The goal of the TPNW was to delegitimise nuclear weapons from humanitarian perspective. (Egeland et al. 2018)

The non-proliferation regime presents a critical effort in maintaining global security. While this regime faced many challenges it continues to be a vital framework for international security.

Nuclear Non-prohibition Regime

Nuclear prohibition regimes are a vital part in efforts to limit the proliferation and use of nuclear weapons. These regimes have been reflecting the global concerns and power dynamics and have significantly evolved though a number of treaties and agreements reflecting the destructive potential of nuclear warfare. This overview traces the important historic developments of those regimes.

Bombings of Hiroshima and Nagasaki marked the dawn of the nuclear age in 1945, showing the devastating effects of nuclear weapons to the general public. This is to show that at first nuclear weapons were viewed as a legitimate weapon of war, it had represented the initial stage (the first stage of Nadelmann's regime creation). The 1950s and 60s represented the first strides for nuclear prohibition regimes, first attempts to control or even eliminate such weapons. Understanding the power of nuclear weapons that at that time created a dichotomy in the US military strategies. First, nuclear weapons had been increasingly viewed as a dedicated class of weaponry, represented by Harry Truman's[5] statements, in the face of the fear and the destructive power of nuclear weapons. Second, nuclear weapons — despite their public perspective — were deemed as an integral part of the US security and have been integrated within military policies. Therefore, the emergent dichotomy, on one hand seeing nuclear weapons as something that should not be used and at the same time, as an integral part of the US security (Tannenwald 2007; Adler 1992; Jervis, Smetana, and Hynek 2015).

The year 1946 marked first attempts of creating a regime for the new nuclear technology. The UN Atomic Energy Commission (UNAEC) was established aiming at promoting peaceful use of nuclear technology, while preventing weaponization. Four main points were raised: the peaceful use of nuclear energy, the possibility to eliminate nuclear

[5] See Harry Truman. 1945. 'Truman Statement on Hiroshima'. https://ahf.nuclearmuseum.org/ahf/key-documents/truman-statement-hiroshima/.

weapons from state arsenals, and putting up effective safeguards for inspections (UN General Assembly 1946).

However, geopolitical tensions, particularly between the United States and the Soviet Union hindered such efforts. Despite this, the competition between the two superpowers brought a common goal: an effort to stop other nations from acquiring nuclear weapons, particularly West Germany and China as Jarvis (2015) remarks.

The Nuclear Non-Proliferation Treaty (NPT)

A historic development in nuclear prohibition was the creation of NPT Treaty which was adopted in 1968 to come into force later in 1970. There are three core ideas in the NPT. The first is embodied in Article II (1975). The non-proliferation of nuclear weapons, non-nuclear-weapons states are obligated not to pursue or receive nuclear weapons. Second, disarmament, although Jarvis (2015) concludes that at first, it was not taken with such priority as it was later in 1980s. Still, it was an important achievement in nuclear prohibition regimes. Although NPT has been praised as the cornerstone of nuclear prohibition regimes, with 191 signatories, due to many challenges, it is continually contested and proven to be difficult to fulfil. Furthermore, the emergence of new nuclear powers (e.g. Pakistan, India, and Israel — none are signatories) hinders the non-proliferation efforts (Jervis, Smetana, and Hynek 2015; Borrie, Caughley, and Hugo 2016). With the more recent development, one could observe the reemergence of arms-race dynamics and the possibility of nuclear weapon use entering the public debate. Wilfred and Chernavskikh (2023) point out that often the significant obstacle in implementing NPT is not on the technical side but rather geopolitics that play a significant role. Moreover, the growing competition in today's multipolar world and shifts in the distribution of power (mainly rising revisionist states) are challenging the effectiveness of the NPT treaty, unlike the Cold War era's bipolar dynamics that once supported its success (Gibbons and Herzog 2022).

Further Prohibition Efforts and Regional Initiatives

One of the recent, no less ambitious efforts in nuclear prohibition is the Treaty on the Prohibition of Nuclear Weapons — TPNW. It was adopted by the UN in 2017, came into force in 2021 (United Nations 2021). It is a binding document that aims at comprehensive prohibition of use, threat of use, storage, testing, production etc. of nuclear weapons. It

has been primarily driven by growing global humanitarian concerns for nuclear-free world. As per previous chapters, this treaty has been driven by humanitarian concerns and the desire for nuclear weapons-free world. One of the issues is that no states that pose nuclear weapons or are protected by nuclear umbrella take part in this treaty, which limits its reach (Wan and Chernavskikh 2023; Patton, Philippe, and Mian 2019; Hajnoczi 2020; Ritchie and Kmentt 2021).

Patton et al. (2019) furthermore addresses the question of the importance of designated international authority that oversees the implementation and adherence to the TPNW treaty. It is suggested that although there are many technical challenges in controlling the adherence to the core disarmament regime, already created international organisations like IAEA should support disarmament verification as a higher international authority. Furthermore, Patton et al. (2019) identify that transnational organisation built gradually though institution building is an efficient way of controlling the adherence to the regime.

The Anti-Ballistic Missile Treaty (ABM)

The ABM treaty was the pinnacle of Cold War strategic arms control agreements. The important notion of this treaty was the establishment of deterrence as the core implementation of nuclear weapons into military strategy, meaning that non-use (defensive use) of nuclear weapons was a main use in US/Soviet military doctrines. Tannenwald (2007) argues that although the purpose of the treaty was the ban of ABMs, the effect of this treaty was a codification of deterrence (i.e. the defensive use of nuclear weapons as their primary objective). The reason the USA has withdrawn from the treaty is noteworthy, as it was an exception among treaties, since the USA opted out because of different external reasons (Kubbig 2005; Wan and Chernavskikh 2023; Tannenwald 2007).

Strategic Arms Limitations Talks

Strategic Arms Limitations Talks (SALT I. and II.) were a series of negotiations between the United States and the Soviet Union during the Cold War aiming at curbing the ongoing arms race between the two superpowers. One can observe that the reasons behind the negotiations were of strategic nature. Recognising that the pursuit of increasing the number of nuclear weapons was unstable, development of SALT I in 1972 represented a mutual recognition of deterrence based on parity and mutual vulnerability. This treaty effectively limited the capacity for land- and submarine-based missiles, and further

modernization could be done only by replacing older systems. Together with the ABM treaty, those actions helped stabilise the situation. It is also noteworthy that SALT I was accompanied by further deliberations about crisis communication which were possible to be based around the presumption of non-offensive use of nuclear weapons (Tannenwald 2007; Bresler 1982; Wan and Chernavskikh 2023).

The subsequent success of SALT I led to more ambitious SALT II treaty. Building on the initial limitation, deeper cuts into the total number of nuclear weapons were made. Salt II was more thorough in some respects than SALT I. It has provided both countries with more robust system of verification one's commitments, there was an exchange of information about number and types of strategic offensive arms. Despite the fact that SALT II was not ratified by the US Senate, both countries initially complied with the treaty, demonstrating common strategical interests in arms control. The continued adherence demonstrated the mutual benefit of limiting strategic arsenals and maintaining parity (Fryer and Levengood 1979; Bresler 1982; Tannenwald 2007; U.S. Department of State, n.d.).

START Treaties

The Strategic Arms Reduction Treaties are a series of agreements aiming at reducing the number of strategic weapons. Those treaties followed the trend of international politics during the post-Cold War era, addressing the need to manage nuclear arsenal and to address the decline in deployed weapons. For the first time, both Russia and the United States agreed to actually reduce the number of delivery devices for nuclear weapons as opposed to only capping their growth. START I was the most comprehensive treaty until that time, apart from lowering the number of weapons, both parties agreed upon verification and on-site inspections. Furthermore, after the dissolution of the Soviet Union, Russia continued with its pledges and Ukraine, Belarus, and Kazakhstan became members of NPT treaty and were about to become non-nuclear members (Thomson 1999; Wan and Chernavskikh 2023; Tannenwald 2007; Rogers, Korda, and Kristensen 2022; Schenck and Youmans 2011).

Start II, which was supposed to be a continuation of START I treaty, have never came into force. Start II was at first blocked from the US side, to be later opposed in Russia as well. Russian opposition stemmed partially from concerns about the US missile defence advancements at the brink of the century. Although Start II has never entered into force,

further reductions in nuclear-capable delivery systems were negotiated, it had a significant effect on both Russian and US strategic planning (Rogers, Korda, and Kristensen 2022; Schenck and Youmans 2011).

New START treaty lowers the number of nuclear warheads by three quarters compared to the original START treaty. In 2010, it replaced the START I treaty that expired in 2009. Crucially, New START includes provisions for on-site inspections, yet they are not as extensive as the previous START treaty was in some respects (Schenck and Youmans 2011). Furthermore, the treaty limited the number of undeployed nuclear-capable delivery devices which presented the shortcoming of previous treaties (Gottemoeller 2021).

"Most importantly, however, the New START Treaty created what has often been referred to as the 'gold standard' of verification: procedures governing the conversion and elimination of strategic offensive arms, the establishment and operation of a database of treaty required information, transparency measures, a commitment not to interfere with national technical means of verification, the exchange of telemetric information, the conduct of on-site inspection activities, and the operation of the Bilateral Consultative Commission." (Rogers, Korda, and Kristensen 2022, 349)

The literature overview demonstrated that there are various schools of thought that explain regime theory and subsequent prohibition regimes. There are three waves of thought characterising regime theory. First, neoliberalism and neorealism, second, cognitivism, and third, radical constructivism/post-structuralism. All three waves demonstrated the shift from state-centric thinking to more in-depth examination of regimes, recognizing the role of non-governmental agencies, social phenomena, like taboos and stigmatization. Recognizing the influence of non-governmental societal structures that influence regime creation both in inner state politics and on an international level. The nuclear prohibition regimes revealed that in most cases, the focus is on the numerical reductions of nuclear weapons and delivery devices supervision. This might be due to the close connection between nuclear weapons and nuclear states' defence strategies, i.e. nuclear deterrence. The current conflict between Ukraine and Russia brought the talk about nuclear weapons into the public debate and highlighted the moral

conundrum of using nuclear weapons as a deterrent on one side and at the same time, not accepting the use of nuclear weapons on the other[6].

[6] On the current Rusian and European public attitudes towards the use of nuclear weapons see Smetana, Michal, and Michal Onderco. 2023. 'From Moscow with a Mushroom Cloud? Russian Public Attitudes to the Use of Nuclear Weapons in a Conflict With NATO'. Journal of Conflict Resolution 67 (2–3): 183–209. https://doi.org/10.1177/00220027221118815. Or Onderco, Michal, Michal Smetana, and Tom W. Etienne. 2023. 'Hawks in the Making? European Public Views on Nuclear Weapons Post-Ukraine'. Global Policy 14 (2): 305–17. https://doi.org/10.1111/1758-5899.13179

Machine Learning

The rapid progress in machine learning in recent decades has sparked a diverse utilization of ML in many fields, from scientific to commercial to creative. So far, machine learning is proving to be one of the most influential emerging technologies. Together with wide application come into question the regulation and security issues connected to the development and deployment of machine learning solutions. Machine learning, at its core, is a development of algorithms that enable computers to learn from data and make predictions based upon them. In the current decade, the proliferation of ML has been witnessed in many spheres (military, science, commerce), the general public has become more familiar especially with generative AI, however, its negative impacts have been increasingly recognised and proved to be a security risk in many instances (e.g. fuelling disinformation campaigns, campaign on 'killer robots', explosion of generated content).

This section will delve into the issues of ML, therefore, it is salient to understand the theoretical foundations of how ML works, what its pitfalls are, and why it succeeds. This section looks into the essential concepts and principles of ML, providing a relevant review for ML in the field of security area. Firstly delving into the theoretical underpinnings of machine learning and its paradigms and continuing to the key issues of training ML models. This section should provide the reader with sufficient overview of the essential principles to help one navigate the advantages and pitfalls of ML deployment in prohibition and non-proliferation regimes.

Machine learning represents a field of computer science which focuses on learning from given data (algorithms learnt from vast datasets). Based on the data the program looks for patterns and inferences which enable it to predict outcomes based on the given data. Mitchel (1997) defines ML as follows: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"* (Mitchell 1997, 2) Machine learning proves itself especially useful in instances where analysing complex and large datasets where human capabilities are insufficient in a reasonable time frame. Although machine learning research has been carried out during the 20th century (1950s Alan Turing), only the recent development of computation power enabled wider use of ML.

There are several paradigms to machine learning, each is defined by the characteristics of the learning model. There are many models but for the purpose of this thesis only few will be discussed: supervised learning, unsupervised learning, and reinforced learning.

Supervised learning

This is one of the most commonly used form of ML training. The core mechanism is that the algorithm is learning from large dataset that is labelled. In this dataset input-output pairs are known i.e. learning from examples. As an illustration, an image-labelling algorithm is shown pictures of people, houses, and cars, all inside a labelled dataset, from which the output parameters of the algorithm are adjusted to correspond to the labelled pictures in given dataset. Common tasks include classification and regression (Alzubi, Nayyar, and Kumar 2018; Fleuret 2024; LeCun, Bengio, and Hinton 2015). *"These learning methods tend to produce task-specific, specialized systems that are often brittle outside of the narrow domain they have been trained on"* (Bengio, Lecun, and Hinton 2021, 62).

Unsupervised learning

This type deals with unlabelled data, finding patterns and structures within the data. Unlike supervised learning it does not require large, labelled datasets to learn. There is a similarity with human vision that is able to abstract vast amounts of information from unsupervised observing, yet this process is not (Bengio, Lecun, and Hinton 2021; LeCun, Bengio, and Hinton 2015; Alzubi, Nayyar, and Kumar 2018).

Reinforced Learning

Reinforced learning represents the third main ML paradigm alongside supervised and unsupervised learning. Rewards and penalties are the driving, teaching force on an agent that is making decisions in an environment and altering its outputs based on the rewards. Therefore, throughout time, the model should perform better as it is optimizing its expected returns. This is particularly well represented in ML models learning on games (Mahesh 2020; Alzubi, Nayyar, and Kumar 2018).

One of the successes of reinforced learning could be seen in ChatGPT and its progression, particularly form version ChatGPT-3. Reinforced training from human learning has been used to some success, yet some of the major obstacles are the time and cost requirements

to obtain quality human feedback. Furthermore, reinforced learning still requires a lot of repetition to be effective which could be expensive, among other issues (Wu et al. 2023).

Artificial Neural Networks and Deep Learning

Deep learning is a subset of machine learning. Machine learning has been for a long time a dominant form of ML. Conventional techniques limit the processing power of information by the need for a feature extractor which translates the raw data into the subsequent system. To improve on this, an inspiration was taken from human brain. Particularly its function and structure, i.e. networks of neurons with deep layers that are interconnected with variant connection strengths. This has sifted the paradigm of thinking about ML, as previously externally defined parameters (rules) of inference defined the output. On the other hand, deep learning works by changing the strength of connection in the given network, furthermore, the new architecture through neurons and their connections group similar objects together. Deep learning has been enabled by the advances in computing power, as it requires large amount of computation power. It has been empowered by GPUs which, together with the availability of large datasets, enabled the emergence of deep learning. Deep learning utilizes deep neural networks with many layers of interconnected noted. Each layer is optimizing the prediction from one layer on to the next one. The input and output layers are visible while the layers in between are hidden (LeCun, Bengio, and Hinton 2015; Bengio, Lecun, and Hinton 2021; Fleuret 2024; Schmidhuber 2015).

"A standard neural network (NN) consists of many simple, connected processors called neurons, each producing a sequence of real-valued activations. Input neurons get activated through sensors perceiving the environment, other neurons get activated through weighted connections from previously active neurons" (Schmidhuber 2015, 4). Deep learning models are highly complex and there are several different types of neural networks, each with its merits, each suitable for different tasks. This thesis discusses four types of neural networks: Feedforward neural networks (FFN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GAN)

Feedforward neural networks represent the simplest type of a neural network. The main characteristic of this network is the unidirectional data flow – from the input layer through hidden layers to the output layer. This one-directional flow of information is characteristic

and makes feedforward neural network one of the two broader types of artificial neural networks. The other type is recurrent neural networks (Haykin 2009; Egmont-Petersen, de Ridder, and Handels 2002; Cheng and Titterington 1994; Glorot and Bengio 2010).

The key difference in recurrent neural networks is the different flow of information. RNNs are usually used for speech-related tasks. Their characteristic function is the feedback loop. This works in effect as a 'memory', in a given sequence the network keeps all information about the previous elements while processing the next one. This rendered these networks particularly effective in predicting next sequences in a text, i.e. predicting a word in a sentence is in effect a product of probability based on previous words and context, unlike traditional neural networks that treat input and output independently (Haykin 2009; Bengio, Lecun, and Hinton 2021; Schmidhuber 2015; LeCun, Bengio, and Hinton 2015; Fleuret 2024).

Convolutional neural networks (CNNs) are specialized in processing grid-like data, pattern and image recognition. They are able to classify pictures and objects within them, applicable to face recognition as well. CNNs are composed of multiple layers, there is an input and output layer, plus several hidden layers in between. Each node (neuron) is connected to another node with a threshold that needs to be met in order to pass the information from one node to the other (Haykin 2009; Fleuret 2024). *"There are four key ideas behind [CNNs] that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers."* (LeCun, Bengio, and Hinton 2015, 439) There are 3 minimal types of layers in CNN: convolutional layers, polling layers, and fully connected layers. First layer extracts features organised into feature maps that are connected to the previous layer by sets of weights. Localisation of features is there because often in pattern recognition, objects in close proximity are connected to the surrounding ones. The second layer, the pooling layer, groups together similar features, which enables particular neurons to share weights if they belong to the same feature map. The third type, the fully connected layers, combines features to make final predictions, i.e. through the combination of low-level features, it creates higher level features, "*in images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences."* (LeCun, Bengio, and Hinton 2015, 439; Haykin 2009; Egmont-Petersen, de Ridder, and Handels 2002)

Generative Adversarial Networks (GAN) are distinctive neural networks that are used to generate new data resembling a template (e.g. generating a new picture from an existing pictures). This network incorporates two components: a generator and a discriminator. While the generator creates something new, something that differs from the template, the discriminator, which is the adversary, takes a sample from the created object and, by prediction, tries to determine if the generated object is part of the template sample. When discriminator determines that the new object is not a part of the template's dataset, it penalizes the generator. This creates a loop which continues until the generator succeeds. This adversarial process improves the generator's ability to produce realistic data (I. J. Goodfellow et al. 2014; Fleuret 2024).

Neural networks and deep learning represent a pinnacle of modern ML, offering a powerful tool for solving complex puzzles across various domains. Despite their high maintenance costs (computational power and large training datasets), they prove to be a potent tool with increasing deployment in various aspects of human existence.

The Process of ML Model Deployment

This part will briefly present the necessary steps in the development of ML model solutions. In combination with the previous chapters on the core mechanisms of ML, it will provide a solid foundation for the subsequent chapter on ML issues.

Alzubi et al. (2018, 5–6) summarises six stages of ML model development:

1) Collection and Preparation of Data

2) Feature Selection — eliminating features from the dataset that are not relevant for the given task

3) Choice of Algorithm — selecting the best suited algorithm for the task at hand[7]

4) Selection of Models and Parameters — Most of machine learning algorithms require some initial manual intervention for setting the most appropriate values of various parameters.

[7] For a list of algorithms see Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. 'Machine Learning from Theory to Algorithms: An Overview'. Journal of Physics: Conference Series 1142 (1): 012012. https://doi.org/10.1088/1742-6596/1142/1/012012. (p. 10-11)

5) Training — After selecting the appropriate algorithm and suitable parameter values, the model needs to be trained using a part of the dataset as training data.

6) Performance Evaluation — Before real-time implementation of the system, the model must be tested against unseen data to evaluate how much has been learnt using various performance parameters like accuracy, precision and recall.

Once the model is tested and achieves the desired performance, it is ready for deployment, which encompasses several steps to ensure the reliability of the model. Proper deployment is a crucial step for leveraging the potential of the ML model, especially in various domains.

There are 3 main steps in ML deployment. The first is integration into the environment. This entails two steps. First on the hardware side: preparation of the physical infrastructure. This could be done both on site or based on a cloud service. Although, the first step might prove to be resource-demanding, the second step is even more crucial and problematic. This issue often entails both researchers and software engineers with overlapping areas of responsibilities. There are several issues that might arise. From the environmental problems (difference in development/testing/production environments) and adversarial training (especially salient when deployed for use by general public or in hostile environment) to model transparency — the demand for explanation behind the model's predictions (Paleyes, Urma, and Lawrence 2022; Bhatt et al. 2020).

Given the early stages of ML development monitoring, the second step is a vital part of keeping ML systems operational. Monitoring a model's performance is on the one hand essential to ensure it predicts as expected, on the other hand, monitoring on itself proves to be an open problem (Paleyes, Urma, and Lawrence 2022; Bhatt et al. 2020). Accuracy, latency, resource utilization, and drifts in distribution are some of the key metrics. Live ML systems can also run into problems of unintentional feedback loops (Sculley et al. 2015). Furthermore, ML systems run into problems with outliers, ML models could either perform poorly when predicting outside training distribution or become overconfident. One has to acknowledge the importance of balancing the monitoring of ML system while retaining sufficient model performance (Klaise et al. 2020).

Updating is the third main issue of ML system deployment. As with much of the software, ML systems need to be updated so they reflect recent changes in data and environment.

Because environments in which ML models are deployed change, the models need to adapt (Quiñonero-Candela 2009).

In conclusion, the deployment of a ML model is a multifaceted process that requires careful planning and adaptability. The three steps do not represent an exhaustive list but describe the most salient issues and illustrate that although training and creating a ML model is a complex task, deployment presents a correspondingly intricate challenge.

Issues Connected to Machine Learning Models

This section will discuss the problems and issues that are stemming from development and deployment of machine learning models. Understanding these challenges is crucial in predicting potential issues with ML. The increasing adoption of ML models in various domains brings about significant safety and security challenges. Ensuring ML models' operational security is critical in minimizing errors and adversarial attacks.

There are several considerations that go into discussing the security of ML models. The basic distinction is what the adversary is targeting plus the cost-benefit analysis of the attacker: it is assumed that the attacker wants to impose maximal damage with minimal costs. The attack on the ML model could be directed on several levels: the model itself (reverse engineering), data used for training (model poisoning), datasets — extraction of sensitive information, adversarial attacks (targeting the input side of the model) Rosenberg at al. (2021) stipulates 3 possible goals of the attacker: *"1) Confidentiality – Acquire private information by querying the machine learning system 2) Integrity – Cause the machine learning system to perform incorrectly for some or all input. 3) Availability – Cause the machine learning system to become unavailable or block regular use of the system."* (2021, 9)

Data Poisoning

The goal of this attack is to skew the output of the model through corrupting the model during the learning phase by altering the learning data. This might take several forms, such as fake data or sensor spoofing attacks. As it was discussed in previous chapters, models are usually updated regularly, therefore such attack is not only limited to the initial stage of training but can occur even during deployment (I. Goodfellow, McDaniel, and Papernot 2018; Koh, Steinhardt, and Liang 2021). With data poisoning, the adversary could also attempt to create a backdoor access. This is especially problematic if the model

is trained with data that has been scraped from the Internet without inspection (Hendrycks et al. 2022). Data poisoning is hard to discover if one does not have the knowledge about it happening. In the case of image recognition, a viable defence is an image reconstruction. Hu et al. (2023) divides poison attacks into two categories: availability attack, the goal of which is to degrade the ML model, and integrity attack which could be well demonstrated on the backdoor attack.

Adversarial Examples

Targeting the already deployed model, adversarial examples abuse fundamental errors in the network, which enables the adversary to take control over the model, especially its output. This problem is particularly visible in image recognition where adversarial examples could take the shape of small adjustments to a picture, imperceptible by human eye, resulting in recognition failure. Evtimov et al. (2020) write that although adversarial examples, particularly for visual recognition, in the real world deployments are feasible, they prove to be laborious to manufacture, rendering them undesirable in the attacker's view (the cost-benefit analysis). Furthermore, an efficient defence against adversarial examples proves to be both negatively effecting the performance of the model and it has narrow effect. Hence it is rarely used in practice and constitutes an omnipresent phenomenon (Zou et al. 2023).

Reverse Engineering

Adversaries can reverse-engineer trained models through model extraction attacks, where they query the model and use its outputs as training data to train an adversarial model. Therefore, for adversaries, access only to input and output is necessary, not particular knowledge about the model's architecture or its original training datasets. By recreating a model based on outputs of the original, adversaries could create adversarial examples on the reverse-engineered model. Given its sufficient accuracy the adversarial examples would be applicable to the original ML model (I. Goodfellow, McDaniel, and Papernot 2018; Rosenberg et al. 2021). Papernot and McDaniel (2016) demonstrated this successfully as a potent method in attacks against Amazon and Google.

Extraction of Sensitive Information

This is particularly problematic in the field of facial recognition and medical uses of ML models. The question of privacy of ML models proceeds in two directions. First, when

users might not be trusted, the ML model provider needs to prohibit users from accessing sensitive information through queries. Second, when the ML provider is not trusted, it is in the user's interest to shield one's privacy from the ML model. Furthermore, the problem of information extraction is twofold. First, ML models might unintentionally reveal, due to errors in architecture, parts of the training data, thus exposing sensitive information. (Papernot et al. 2018; Papernot 2018; I. Goodfellow, McDaniel, and Papernot 2018; Solaiman et al. 2024) As an example, one might consider medical records, where one could observe how originally anonymized medical records could be attributed to individuals. GANs data synthetisation might prove to be a solution to data anonymization in this case, though (Zhang et al. 2022; Solaiman et al. 2024; Kaissis et al. 2021).

When one is thinking about ML security, it is necessary to consider several angles. The goals of the adversary and possible levels of attack should be considered. The last piece of the puzzle is connected to the cost benefit analysis, it is the level of the adversary's familiarity with the ML model. Rosenberg (2021) summarises 4 kinds of attacks based on familiarity with the model: Black-Box, Grey Box, White Box, and Transparent Box.[8]

"1) Black-Box attack requires no knowledge about the model beyond the ability to query (i.e. input-output access) 2) Gray-Box requires limited degree of knowledge about the targeted classifier […] 3) White-Box attack – the adversary has knowledge about the model architecture and even the hyperparameters used to train the model. 4) Transparent-Box attack – In this case, the adversary has complete knowledge about the system, including both white-box knowledge and knowledge about the defence methods used by defender" (Rosenberg et al. 2021, 10).

The safety and security of ML models are critical aspects that must be addressed to illuminate issues of misuse and malicious attacks. By understanding the various challenges, one could better prepare and evaluate the potential deployment of ML model. Although the field of ML continues to advance rapidly, both in adversarial strategies and

[8] For a comprehensive list of challenges with deploying ML models see Paleyes, Andrei, Raoul-Gabriel Urma, and Neil D. Lawrence. 2022. 'Challenges in Deploying Machine Learning: A Survey of Case Studies'. ACM Comput. Surv. 55 (6): 114:1-114:29. https://doi.org/10.1145/3533378. Furthermore, tradition list of ML related threats see: Hu, Yupeng, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2023. 'Artificial Intelligence Security: Threats and Countermeasures'. ACM Computing Surveys 55 (1): 1–36. https://doi.org/10.1145/3487890. For cyber security related issues with ML see: Rosenberg, Ihai, Asaf Shabtai, Yuval Elovici, and Lior Rokach. 2021. 'Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain'. arXiv. https://doi.org/10.48550/arXiv.2007.02407.

their mitigation, many practices discussed in this section are core mechanics and processes that should change minimally (e.g. the cost-benefit analysis of the malicious actor, Rosenberg's three attacker goals).

Theoretical Framework

As was established in the literature review about regime theory, one could observe a shift in the research, the increasing involvement of non-governmental institutions. Therefore, it is crucial to choose an IR theory that reflects the increasing involvement of NGOs. This will be also pertinent to the ML part of this thesis because one can observe that major ML models are developed by private actors and it could be speculated that further deployment of ML models, even in the field of international regimes, will be heavily dependent of non-governmental actors. For that reason, constructivism (part of the 2nd wave of Hynek's regime theorization) has been chosen, as it allows us to consider both non-governmental actors, international norms and state involvement. This thesis discusses nuclear and cluster munition international regimes with which concepts of humanitarian disarmament, stigmatization and taboos are strongly connected.

Constructivism

Constructivism, among the other prominent theories in security studies, like realism and liberalism, explains the issues of nationality and security instead of state-centralist view through material factors like power, through the lenses of ideas and interests. Hence the central idea is that security is constructed — a product of society (Buzan and Hansen 2009; Williams 2008; Wendt 1995; 1992). Goldstein and Keohane (1993) write that although ideas are the main concept, the principle of transferring ideas into policies is the main issue area. Goldstein and Keohane (1993) postulate three types of beliefs:

1) Basic 'ideas' that define reality — basic concepts that are connected to one's culture and influence one's understanding of reality.

2) 'Principled beliefs' which are normative ideas (right vs. wrong). Those beliefs are instrumental in creating, in case of this thesis the prohibition and non-proliferation regimes. Their fluctuations spark changes in political order (e.g. establishment of human rights after World War 2, bans on cluster munitions, antipersonnel mines, chemical weapons etc.). This corresponds to Nadelmann's second stage of regime creation, the shift

in narratives. Principled beliefs *"translate fundamental doctrines into guidance for contemporary human action"* (Goldstein and Keohane 1993, 9).

3) 'Causal beliefs' which are beliefs about causal effects. They are based on scientific knowledge and serve as a guideline for implementation of established principal beliefs (e.g. the application of non-violence in conflict by Sharp[9], Chenoweth and Cunningham[10]). Causal beliefs reflect the evolution of science and are, therefore prone to more frequent changes than principled beliefs. Causal beliefs are embodied by P. Haas's epistemic communities.

Additionally, both Keohane (1993) and Wendt (1995) agree on the fact that once beliefs are transformed into norms, they could become a part of institution, this creates a stronger and longer-lasting institutionalized belief which strongly influences the regime in which it operates. This is essential to later understand that those institutionalized beliefs are used in regime creation (collective action) in bargaining where persuasion, rather than coercion, creates the need for reasoning, explaining one's actions and positions.

The notion of how institutionalizes beliefs could help in bargaining is a salient point which is connected to game theory and cooperation. *"Social construction is like game theory talk: analytically neutral between conflict and cooperation"* (Wendt 1995, 76). Axelrod (1984) describes the 'Tit for Tat' strategy in game theory (particularly applied to prisoner's dilemma) as one that is particularly successful. He attributes its success to three key features: being nice, retaliatory, and forgiving. Nice — not being the first to defect; retaliatory — responding as soon as possible to defection from the other player (in equal proportion = defection); forgiving — after retaliatory measures return to status quo (not holding grudge). Furthermore, Axelrod (1984) stipulates about emerging cooperation that rationality of the actors is not necessary; nor is trust, nor central authority, as long as one follows the three key features of Tit for Tat strategy cooperation emerges that is based on reciprocity. Although constructivists differ on the analysis of the cooperation, i.e. examining how cooperation (even if for egoist reasons) affect beliefs and norms which in theory shape the cooperation from initial utilitarian to one of shared norms and

[9] Sharp, Gene. 2008. From Dictatorship to Democracy: A Conceptual Framework for Liberation. 3rd U.S. ed. East Boston, MA: Albert Einstein Institution.
[10] Chenoweth, Erica, and Kathleen Gallagher Cunningham. 2013. 'Understanding Nonviolent Resistance: An Introduction'. Journal of Peace Research 50 (3): 271–76. https://doi.org/10.1177/0022343313480381.

commitments (Wendt 1992). The issue of retaliation in such case would also be observed as a function of epistemic communities and state institution, which makes it in turn predictable — basing in on the norms, beliefs adopted by the given state (Goldstein and Keohane 1993).

Although constructivist agree on the basis that norms, identity, and ideas are the key concepts in understanding security. There are conventional and critical constructivists that diverge on explaining changes and legitimacy of norms. Classical constructivist, like the aforementioned A. Wendt, postulate that norms (ideas) are quite stable and the relationship between security and identity is dependent on cultural and historical context. They also work within both realists' and liberalists' assumptions and framework (Buzan and Hansen 2009; Williams 2008). Importantly, Wendt (1992) still recognizes states as one of the dominant players in the international system. Critical constructivism on the other hand sees norms as volatile, dependent on constant contestation. Identity serves as a central concept in finding definition of us versus them. Examination of different competing identities which crystallises into a national identity. Political scientists are less concerned about describing the phenomenon from outside and rather interpreting from within (Williams 2008; Buzan and Hansen 2009).

It is difficult to draw precise lines between constructivism, critical constructivism, and Poststructuralism, as there are many overlapping areas of interest (Buzan and Hansen 2009). Structuralists also comment on the state of affairs in anarchical state and contest Walzian's view that states exist in a state of anarchy which determines the structure of the international system. Went (1992) argues that (with reference to game theory) despite absent higher authority states, cooperation will appear almost inevitably because of instrumental reciprocity. Therefore, anarchy represents only a temporary state of affairs which could be changed through changing norms and cooperation, i.e. through norms states could construct the understanding of affairs and one's security.

"While constructivists agree that security is a social construction, however, attempts to point more explicitly to how security works and how we might study its construction are more controversial. Most constructivists have avoided this question, but the Copenhagen School has attempted to develop a more coherent theory […]" (Williams 2008, 67).

This thesis primary utilises classical constructivism (Wendt, Keohane) with added observations from game theory by Axelrod as a theoretical framework to explore the utilization of machine learning in prohibition and non-proliferation regimes. Constructivism should allow one to connect the technical specifications of ML models with regime theory. However, there are still some concepts that need to be addressed. Norms play important role both in regime theory and constructivism. Although constructivism does not shed much light on trust and legitimacy of NGOs, international institutions, and other actors, the mere presence of the phenomenon is sufficient for this thesis. Furthermore, it is important to recognise that states — as actors in the international arena — hold the power and cannot, apart from coercion by other state, be forced to act against their will.

Humanitarian Disarmament

Humanitarian disarmament is based on the principle of reduction of human, particularly civilian, suffering caused by conflict and violence. Unlike traditional disarmament where disarmament efforts are fuelled by state interests, security, and strategic aims, humanitarian disarmament prioritises the protection of civilians and the mitigation/minimalization of human suffering. Efforts in this area represent an intersection of human security and disarmament. Humanitarian efforts span across several issue areas, such as antipersonnel mines, arms trade, cluster munition, killer robots, nuclear weapons, etc. Its origins could be traced to the 1997 Mine Ban Treaty (Ottawa Treaty) (Docherty 2018; Borrie and Randin 2006). Further significant development is tied to the Conference on Cluster Munitions which used the phrase ‘unacceptable harm’ (First used in Oslo Declaration in 2007) (United Nations 2008).

Both Convention on Cluster Munitions and Mine Ban Treaty are framed as a humanitarian issue, the argumentation is based around moral and ethical considerations. This represents a shift from state-centric disarmament to focus on people and civilians in particular. Such a turn also highlighted the increasing influence of civil society in the process of prohibition regime creation[11]. The subsequent treaty on cluster munitions signified a more

[11] The case of Land Mine Treaty the participation of civil society see: Williams, Jody, Stephen D. Goose, and Mary Wareham, eds. 2008. Banning Landmines: Disarmament, Citizen Diplomacy, and Human Security. Lanham: Rowman & Littlefield.

permanent shift and humanitarian disarmament was well established by 2008 (Finaud 2017; Garcia 2015).

Docherty (2018) identifies three main humanitarian provisions that characterise humanitarian disarmament treaties in general: absolute preventive obligation, remedial measures, and cooperative approaches to implementation. First, absolute preventive obligation has usually several parts, it is an absolute and wide-ranging provision; additionally it covers provisions towards destruction of any already existing material. Second, remedial measures often build on previous obligations with 'positive obligations' which could be ranging from cleaning contaminated areas to providing education and healthcare to affected populus. Those provisions are to remedy past harm and possible future harm as well. Third, cooperative approaches to implementation (building on previous provisions) suggest an international cooperation in order to enable effected states to meet their respective obligations.

With the emerge of humanitarian discourse and language, one could incorporate it with the rest of the theory. If one applies the previous constructivism in regard to regime theory, one could acknowledge the diffusion of humanitarian norms or beliefs per Wendt's terminology and recognise the emergence of humanitarian disarmament prohibition and non-proliferation regime. Furthermore, if one examines the conditions of humanitarian regime creation by Garica (2015, 62) in reference to prohibition regime creation[12], it follows the same model.

Stigmatization and Taboo

There is one last puzzle that could be addressed further: the case of how the shift in narratives in regime creation emerges. It could help with explaining why nations that do not have or want to pursue given weapons under prohibition and non-proliferation regime take part in those regimes. In her book, Tannenwald (2005) looks at the case of nuclear taboo. Tannenwald explains nuclear taboo in the broader perspective of narratives and the role of anti-nuclear movements and anti-nuclear states. Nuclear taboo managed to overcome the resistance from nuclear states and in many instances helped the prohibition regime (Tannenwald 2005; Schelling 2007; F. Sauer 2016).

[12] Nadelmann's 4 stages of regime creation

There are several strategies that actors could employ in their campaign towards stigmatization, such as mobilization, social pressure, lobbying, and coalition building. This demonstrates the importance of non-state actors, like NOGs and civil society, which could be activated via awareness-raising, naming, and shaming. This increases the societal pressure on governments to adhere to the desired narrative (Rosert and Sauer 2021).

It was established by constructivism that norms play an important role in regime creation and that influencing narratives affects policies. Price (1995) argues that norms played a significant role in the non-use of chemical weapons during World War II. Those principles are well established and are used in current campaigns on killer robots, blinding lasers, and cluster munitions[13] (Rosert and Sauer 2021).

Sauer and Reveraert (2018) further discus the importance of normative change as reframing nuclear weapons in the TPNW treaty from appreciated, prestigious weapons to stigmatised undesirable weapons. Although, as was established in the chapter on constructivism, normative changes, as in the principled beliefs[14], are rigid, hard to change, and their alternation is a prerequisite for political change and subsequent regime creation.

Stigmatization proved to be a valuable asset in disarmament campaigns in several cases, like nuclear disarmament or cluster munition. Framing these weapons as morally and ethically acceptable produces a framework for prohibition and non-proliferation regime creation. Mobilizing public and non-governmental organisations creates societal pressure on governments and helps to shift public narratives, which is one of the key factors in regime creation.

Chapter Summary

The core of this thesis relies predominantly on constructivism and its view on regime creation. From this theoretical standpoint, one could include the actions of non-governmental organisations (particularly with regards to Nadelmann's stages of regime creation) and normative practices associated with humanitarian disarmament, in particular Goldstein and Keohane's classification of beliefs and the role of epistemic

[13] The example of public pressure as one of the reasons behind stopping transfers of cluster munitions to Saudi Arabia see: Wareham, Marry. 2016. 'U.S. Must Stop Giving Cluster Munitions to Saudis'. HuffPost. 6 June 2016. https://www.huffpost.com/entry/on-cluster-munitions-a-te_b_10319504.

[14] Goldstein and Keohane's three types of beliefs

communities. This classification allows for connection with game theory and Axelrod's three features of successful strategies for successful cooperation. The aggregation of these theoretical standpoints permits the examination of theoretical areas where machine learning could be utilized for its possible contributions on either regime creation or regime adherence control. The theoretical examination of these potential places for utilization of ML is the goal of this thesis. Furthermore, the theoretical underpinning of ML and the understanding of potential issues with ML deployment should further enhance the accuracy of this scrutiny revealing potential problems with ML deployment that would have been missed if one only relied on regime theory).

Humanitarian disarmament represents a critical approach in international relations, emphasizing the protection of civilians and the reduction of human suffering. By focusing on the humanitarian imperative, norm-creation and promotion of comprehensive strategies for prevention and remediation of impacts on mainly civilian population, Humanitarianism provides a theoretical platform for further integration of ML solutions into regime theory.

Findings and Discussion

From the IR theoretical perspective, the most suited strategy of ML deployment in prohibition regimes has several key attributes. First, such deployment will increase the knowledge of the actor about the regime, increasing his influence on the regime. Second, it is the ML deployment that enables actors to quickly recognise behaviour in violation of the regime. This facilitates options to instate a retaliatory action in response (e.g. name and shame campaigns, diplomatic measures) in accordance with the logic of Axelrod's game theory[15]. In combination, the acquired knowledge further increases actors' ability (utilizing moral and emotional factors) to influence norms and their creation (i.e. increasing the impact on the ability to successfully shift narratives/change principled believes).

Research Findings

Cluster Munitions

Current enforcement regime

Cluster munitions have been the focus of significant international efforts aimed at prohibition and elimination. Widespread public concern with CM, their indiscriminate nature and long-lasting effects led to the creation of the Cluster Munition Coalition[16] (CMC) in 2003, which effectively started the establishment of the prohibition regime (Docherty 2007; Bolton and Nash 2010; Human Rights Watch 2010b). The primary mechanism for enforcement is the Convention on Cluster Munitions (CCM) which provides a legal framework.

Although the military utility of cluster munitions has been reported to be declining (Docherty 2007), the continued use of these munitions in conflicts and their retention by major world powers suggests that they still have military utility. This in fact puts even bigger stress on the role of public society, NGOs, and middle powers to wider the regime.

---

[15] Quick and proportional reaction creates predictable behaviour, and it aids cooperation.

16 "Launched in The Hague in November 2003 by 85 NGOs from nearly 50 countries, the CMC aims to provide a coordinated global civil society response to the numerous problems created by cluster munitions." (Cluster Munition Coalition 2023)

CMC (2023) reports that the most recent use has been reported in Ukraine, Myanmar, and Syria.

There are two layers of legal obligations. The first are the obligations that states agreed to by participating in the 2008 Convention on Cluster Munitions. The second are state laws, which are adopted following Article 9 of CCM, which stipulates that: *"Each State Party shall take all appropriate legal, administrative and other measures to implement this Convention, including the imposition of penal sanctions to prevent and suppress any activity prohibited to a State Party under this Convention undertaken by persons or on territory under its jurisdiction or control"* (United Nations 2008, 17).

CMC (2008) registers 33 states that have so far enacted state legislation in accordance with CCM's provisions. It is well established that NGOs and middle powers had a profound positive effect on the prohibition regime (Bolton and Nash 2010; Rappert and Moyes 2009; Docherty 2007).

States are obligated by the CCM to submit a first report on the state of cluster munitions and then to do yearly reports. Yet CMC (2023) states that only a fraction of states actually provides those reports regularly.

Machine Learning Deployment

To better structure the potential uses of ML in the cluster munitions prohibition regime, the analysis will be divided into several parts following the structure of the CCM: prohibitions (of use, stockpiling, production and development, and transfer) and obligations (destruction of stockpiles of cluster munitions, clearance and destruction of cluster munition remnants, and assistance to victims and survivors).

1) Prohibitions of use

In the case of cluster munitions, monitoring the use of CM is one of the key areas which helped to shape and create the regime (main source of stigmatization). Monitoring the use of ML provides actors with the biggest leverage in bargaining for prohibition. The ability to capture the use of cluster munitions, therefore, proves itself to be one the key elements of the whole regime.

There are several utilisations of ML. First, image recognition could prove very useful in recognising the use of cluster munitions in images. Such a system could automatically

detect debris (used CM casings and parts) in vast amounts of imagery, significantly reducing the time and effort required for manual analysis. Such efforts at detecting cluster munitions utilizing computer vision have been successfully tested by Harvey and LeBrun (2023) both in photographic and video images[17]. Although this project specialised in cluster munitions, there are other research projects that focus more generally on unexploded ordnance (Craioveanu and Stamatescu 2024). Such efforts make the recognition of negative cluster munitions effects viable. Both models are based on convolutional neural networks.

The two main possible problems with deploying ML models in mapping the use of cluster munitions are adversarial examples and data poisoning. In the case of adversarial examples, the attacker might flood the resource materials (provided that data for analysis is scraped from open sources, primarily internet and social media) with data that will be falsely attributed (putting the model's reliability into question). Data poisoning can compromise the model's integrity, but learning from controlled, rendered data (as in the case of VFrame project) during the initial training phase can mitigate the issue.

2) Prohibition of Production, Development, and Stockpiling

Given that control of production and development mainly relies upon annual self-reporting on compliance by states themselves, efforts in deploying ML solutions might prove to be challenging and ineffective, or not cost-effective.

Cluster munitions can be identified by the unique craters they create when tested. Duncan et al. (2023) achieved some success in mapping artillery craters, although their method focuses on detecting the crates rather than analysing their shapes and characteristics. This approach is limited and depends on satellite or aerial imagery, complicating its effectiveness.

Despite the fact that data are scarcer, such characteristic (controlled input) could prove to be advantageous in easier recognition of reliable sources for input material, making the potential model much safer.

3) Prohibition of Transfer

[17] For more about synthetic datasets (utilizing both rendered and printed data) see: https://vframe.io/3d-rendered-data/

This is another area where the control of transfer is reliant of self-reporting (Convention on Cluster Munitions and Geneva International Centre for Humanitarian Demining 2016; Cluster Munition Coalition 2023). Furthermore, using publicly available data and OSINT methods fused with ML could prove cumbersome and cost-ineffective. Therefore, monitoring transfer prohibition, although vital part of the regime is not an ideal space for ML deployment.

4) Obligation to Destroy Stockpiles of CM

Cluster munitions are difficult and expensive to destroy, often involving their reverse engineering because of their initial design characteristics. Because it is not possible to reliably establish the number of cluster munitions among both member and non-member states of CCM, it makes externally monitoring their destruction futile and reliant on self-reporting, rendering ML deployment vain.

5) Obligations to Clear and Destroy CM Remnants

This point is closely connected to the first point about monitoring the use of CM. Because image recognition enables one to monitor the uses of CM. One could record the contaminated areas and mark them to be cleared. To further assist, ground-based detection systems can be used in tandem with ML, such as metal detectors and ground-penetrating radars. The main advantage of these avenues is that they rely on other methods of detecting areas for demining, which is crucial. Because the time difference between use and subsequent demining efforts might hamper image-based recognition by creating a distributional shift in the environment.

6) Obligations Towards Victims and Survivors

Given the multifaceted nature of this point (provisions touching on medical care, psychological support, social and economic inclusion), implementing ML solutions might be possible for individual issue areas, but given the scope of the problem, it falls outside the possibilities of this thesis. Moreover, it does not directly relate to the immediate issues of cluster munitions monitoring.

Nuclear Weapons

Currently, the use of nuclear weapons is operating mainly under NPT and TPNW treaties. Given the special place that nuclear weapons occupy in national defence, mainly as

deterrence weapons monitoring is predominantly a state affair where International Atomic Energy Agency (IAEA) plays a role of mediator and enabler of international cooperation in monitoring nuclear weapons. NPT in Article 3 states that each member state is required to conclude a safeguards agreement with the IAEA (IAEA 2016). Furthermore, one has to acknowledge the specificity of this regime in contrast to the regime on cluster munitions — the fact that since World War II, nuclear weapons have not been used in any conflict and to this day states rely only on their deterrent capability, and a strong non-use taboo narrative. Although classical control mechanism through state-led treaties is still primary, due to expanding access to public digital data and technological advancements in data analytics, non-state actors have increasing potential to monitor the regime (The Nuclear Threat Initiative 2014).

Nuclear prohibition regime is aiming at the elimination of the threat posed by nuclear weapons and involves complex and multifaced efforts including compliance, monitoring, verification, and enforcement. ML with its capability to analyse vast amounts of data, and pattern recognition can enhance disarmament and compliance efforts. This section explores the potential application of ML in several issue areas.

Given the scope of this issue area and various actors that participate in the nuclear prohibition regime, the possibility of ML deployment in this regime has been, for the sake of convenience, divided based on the life cycle of a nuclear weapon: 1) Material production 2) Warhead assembly and deployment 3) Storage 4) Dismantlement.

1) Material Production

The crucial part of early stages of nuclear weapons development. There are several areas where ML could be deployed. From the earliest stages leveraging satellite imagery with various sensor data to detect irregularities. Grace et al. (2019) use remote sensing to detect uranium mining and milling utilizing various sensor data (thermal, optical, near-infrared, hyperspectral imagery), ML models could be used to assess and quickly detect abnormalities for inspections (Grace Liu et al. 2019; King's College London 2016).

Furthermore, increasing surveillance creates more data that needs to be processed. Which is a tedious process and the increasing technological capabilities to track nuclear material, on one hand, could further increase control and accountability, on the other hand, increasing amounts of data prove to be a challenge for human experts to go though (often

looking for small abnormalities among vast amounts of data). This is an area of great possibility for ML solutions (Mullens et al., n.d.; Revill and Garzón Maceda 2022; The Nuclear Threat Initiative 2014; Keel et al. 2010).

2) Warhead Assembly and Deployment

There are several ways in which warheads are accounted for. The current regime between the United States and Russia has been based on verification of delivery vehicles, making the count of warheads indirect. Although there is data available from open sources (treaties disclosures, governmental statements, declassified documents, and budgetary information), data from satellite imagery and other non-state entities (media reports) help to complete the picture and enable analysts to estimate the numbers of nuclear warheads (Kristensen and Korda 2023; Kristensen et al. 2024; Chen et al. 2016).

Machine learning models could speed up the process of analysis by tracking the movements of missile launchers and other delivery systems and perform a contextual analysis of published documents. Yet, there are many practical problems and considerations, such as tracking delivery devices might in some cases prove futile as in the case of dual-purpose weapons capable of both conventional and nuclear payload.

3) Storage

"As part of its efforts to detect any undeclared nuclear material and activities and to complement other safeguards relevant information, the IAEA collects and analyses a variety of openly available information in a diversity of formats including text-based information photographs, and satellite imagery. In response to changes in the overall information landscape in recent years, it has also strengthened its capabilities to utilize multimedia-based information." (IAEA 2020, 13)

Arterburn, Dumbacher, and Stoutland (2021) used machine learning models to identify abnormalities in trading data greatly assisting in uncovering illicit trade with nuclear materials and equipment.

The increasing accessibility of publicly available information and their volume is an opportunity for the use of machine learning to improve both speed and quality of data analysis.[18]

4) Dismantlement of Nuclear Weapons

This is so far the most challenging part because although *"current verification techniques have allowed for reductions in deployed nuclear forces because both sides can more easily verify limits on the number of delivery vehicles, such as missiles and aircraft"* (Hickey 2022, 94), a dismantlement of nuclear warheads is difficult because of their classified status. Therefore, other than deploying ML models to track nuclear materials as discussed in aforementioned storage paragraph, ML solutions are not useful.

Problems with deploying ML models in nuclear weapons prohibition regime

There is an inherent security problem with deploying ML solutions in a highly sensitive environment. Such models might be vulnerable to extraction of sensitive information and in general attacks from an adversary. This is due to the sensitivity of information and the fact that nuclear weapons occupy a special status in states defence, therefore the actors might be willing to accept higher costs for attacking the system.

Discussion

The analytical part had revealed that ML could in some areas significantly enhance the control of adherence to prohibition regimes by both state and non-state actors.

For cluster munitions the biggest avenue for ML deployment is to help with monitoring the use of cluster munitions. The reason for that is the capacity of ML models to process vast amounts of data. Which enables one to harness the increasingly large amounts of publicly available data. The ability to quickly recognise and analyse the use of cluster munitions puts one into advantageous situation, as one could promptly react and address the transgressions accordingly, following the logic of game theory and constructivism. Using gained knowledge to influence the norms and identities by engaging in retaliatory reaction to adversary who broke the prohibition regime.

[18] For UNIDIR verification experiment of storage facility see: https://unidir.org/menzingen-verification-experiment-2/

Utilisation of machine learning in nuclear weapons prohibition is a more complicated matter because of two systematic issues. First, the special regime of nuclear weapons, being both a subject to non-use taboo and at the same time, being tight to the logic of deterrence. Similarly to the cluster munitions case, machine learning could be potentially deployed by both states and non-state actors to control the adherence to the regime. Yet there are many complications, from the sensitive data that might be at risk of being extracted by adversarial actors to the "high stakes" of nuclear armament which means that protentional adversaries will put more resources against deployed ML models.

To answer the research question, machine learning could significantly help with monitoring prohibition regimes. Its potential lies especially in the field of open source intelligence, where it enables to process the cast amounts of data and streamline the collection of data for further analysis. This could, in turn, significantly boost non-state entities like NGOs and their campaigns which will be able to faster react and hence more effectively contribute to the enforcement of international prohibition regimes.

Therefore, in this thesis it is argued that the utilisation of machine learning, especially by non-governmental organisation, enables them to increasingly influence prohibition regimes by shifting narratives by effectively exposing perpetrator's defection from the regime, and instating proportional retaliation in time.

Furthermore, there is a problem with open-source intelligence and trustworthiness of data gathered from open sources which have not been addressed in this thesis. The main reason is that although it is a quintessential issue area, it is out of the scope of this work. The main priority was to establish theoretical foundations between machine learning and regime theory. However, Revill and Garzón (2022) identify two main issues with data from open sources — data authentication and data corroboration (i.e. revealing if the data are genuine and that it is possible to replicate the processes of data validation).

References

Adler, Emanuel. 1992. ‘The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control’. *International Organization* 46 (1,): 101–45.

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. ‘Machine Learning from Theory to Algorithms: An Overview’. *Journal of Physics: Conference Series* 1142 (1): 012012. https://doi.org/10.1088/1742-6596/1142/1/012012.

Arterburn, Jason, Erin D Dumbacher, and Page O Stoutland. 2021. ‘SIGNALS IN THE NOISE’. *Nuclear Threat Unitiative*.

Axelrod, Robert M. 1984. *The Evolution of Cooperation*. Rev. ed. New York, NY: Basic Books.

Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton. 2021. ‘Deep Learning for AI’. *Communications of the ACM* 64 (7): 58–65. https://doi.org/10.1145/3448250.

Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. ‘Explainable Machine Learning in Deployment’. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–57. Barcelona Spain: ACM. https://doi.org/10.1145/3351095.3375624.

Bolton, Matthew, and Thomas Nash. 2010. ‘The Role of Middle Power–NGO Coalitions in Global Policy: The Case of the Cluster Munitions Ban’. *Global Policy* 1 (2): 172–84. https://doi.org/10.1111/j.1758-5899.2009.00015.x.

Borrie, John, Tim Caughley, and Torbjørn Graff Hugo. 2016. *A Prohibition on Nuclear Weapons: A Guide to the Issues*. International Law and Policy Institute (ILPI).

Borrie, John, and Vanessa Martin Randin. 2006. *Disarmament as Humanitarian Action: From Perspective to Practice*. Geneva, Switzerland: United Nations Institute for Disarmament Research.

Bresler, Robert J. 1982. ‘The Tangled Politics of SALT’. *Arms Control* 3 (1): 3–12. https://doi.org/10.1080/01440388208403738.

Buzan, Barry, and Lene Hansen. 2009. *The Evolution of International Security Studies*. Cambridge, UK ; New York: Cambridge University Press.

Chen, Cliff, Crystal Dale, Sharon DeLand, Angela Waterworth, Doug Keating, and Matthew Oster. 2016. ‘Developing a System Evaluation Methodology for a Warhead Monitoring System’. *Institute of Nuclear Materials*, June.

Cheng, Bing, and D. M. Titterington. 1994. ‘Neural Networks: A Review from a Statistical Perspective’. *Statistical Science* 9 (1): 2–30.

Cluster Munition Coalition. 2023. ‘Cluster Munition Monitor 2023’. *ICBL-CMC*. www.the-monitor.org.

Convention on Cluster Munitions and Geneva International Centre for Humanitarian Demining. 2016. *A Guide to Cluster Munitions*. 3.

Craioveanu, Marian G., and Grigore Stamatescu. 2024. ‘Detection and Identification of Unexploded Ordnance Using a Two-Step Deep Learning Methodology’. In *2024 32nd Mediterranean Conference on Control and Automation (MED)*, 257–62. https://doi.org/10.1109/MED61351.2024.10566207.

Docherty, Bonnie. 2007. ‘The Time Is Now: A Historical Argument for a Cluster Munitions Convention’. *Harvard Human Rights Journal* 20:53–88.

———. 2018. ‘A “Light for All Humanity”: The Treaty on the Prohibition of Nuclear Weapons and the Progress of Humanitarian Disarmament’. *Global Change,*

Peace & Security 30 (2): 163–86. https://doi.org/10.1080/14781158.2018.1472075.

Duncan, Erik C., Sergii Skakun, Ankit Kariryaa, and Alexander V. Prishchepov. 2023. 'Detection and Mapping of Artillery Craters with Very High Spatial Resolution Satellite Imagery and Deep Learning'. *Science of Remote Sensing* 7 (June):100092. https://doi.org/10.1016/j.srs.2023.100092.

Egeland, Kjølv, Torbjørn Graff Hugo, Magnus Løvold, and Gro Nystuen. 2018. 'The Nuclear Weapons Ban Treaty and the Non-Proliferation Regime'. *Medicine, Conflict and Survival* 34 (2): 74–94. https://doi.org/10.1080/13623699.2018.1483878.

Egmont-Petersen, M., D. de Ridder, and H. Handels. 2002. 'Image Processing with Neural Networks—a Review'. *Pattern Recognition* 35 (10): 2279–2301. https://doi.org/10.1016/S0031-3203(01)00178-9.

Evtimov, Ivan, Weidong Cui, Ece Kamar, Emre Kiciman, Tadayoshi Kohno, and Jerry Li. 2020. 'Security and Machine Learning in the Real World'. arXiv. https://doi.org/10.48550/arXiv.2007.07205.

Finaud, Marc. 2017. *'Humanitarian Disarmament': Powerful New Paradigm or Naive Utopia?* Geneva Centre for Security Policy.

Fleuret, François. 2024. *The Little Book of Deep Learning*. https://fleuret.org/public/lbdl.pdf.

Fryer, Keith E., and J. Michael Levengood. 1979. 'Arms Control: SALT II - Executive Agreement or Treaty Recent Development'. *Georgia Journal of International and Comparative Law* 9 (1): 123–36.

Garcia, Denise. 2015. 'Humanitarian Security Regimes'. *International Affairs* 91 (1): 55–75. https://doi.org/10.1111/1468-2346.12186.

Gerring, John. 2017. *Case Study Research: Principles and Practices*. Second edition. Strategies for Social Inquiry. Cambridge: Cambridge University Press.

Getz, Kathleen A. 2006. 'The Effectiveness of Global Prohibition Regimes: Corruption and the Antibribery Convention'. *Business & Society* 45 (3): 254–81. https://doi.org/10.1177/0007650306286738.

Gibbons, Rebecca Davis, and Stephen Herzog. 2022. 'Durable Institution under Fire? The NPT Confronts Emerging Multipolarity'. *Contemporary Security Policy* 43 (1): 50–79. https://doi.org/10.1080/13523260.2021.1998294.

Glorot, Xavier, and Yoshua Bengio. 2010. 'Understanding the Difficulty of Training Deep Feedforward Neural Networks'. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–56. JMLR Workshop and Conference Proceedings. https://proceedings.mlr.press/v9/glorot10a.html.

Goldstein, Judith, and Robert O. Keohane. 1993. 'Ideas and Foreign Policy: An Analitical Framework'. In *Ideas and Foreign Policy: Beliefs, Institutions, and Political Change*, edited by Social Science Research Council (U.S.). Cornell Studies in Political Economy. Ithaca: Cornell University Press.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. 'Generative Adversarial Networks'. arXiv. http://arxiv.org/abs/1406.2661.

Goodfellow, Ian, Patrick McDaniel, and Nicolas Papernot. 2018. 'Making Machine Learning Robust against Adversarial Inputs'. *Commun. ACM* 61 (7): 56–66. https://doi.org/10.1145/3134599.

Gottemoeller, Rose. 2021. *Negotiating the New START Treaty*. Cambria Press.

Grace Liu, Joseph Rodgers, Scott Milne, Margaret Rowland, Ben McIntosh, Mackenzie Best, Octave Lepinard, and Melissa Hanham. 2019. ‘Eyes on U: Opportunities, Challenges, and Limits of Remote Sensing for Monitoring Uranium Mining and Milling’. *James Martin Center for Nonproliferation Studies*, January. https://www.nonproliferation.org/wp-content/uploads/2019/01/op44-eyes-on-u.pdf.

Haas, Peter M. 1992. ‘Introduction: Epistemic Communities and International Policy Coordination’. *International Organization* 46 (1): 1–35.

Hajnoczi, Thomas. 2020. ‘The Relationship between the NPT and the TPNW’. *Journal for Peace and Nuclear Disarmament* 3 (1): 87–91. https://doi.org/10.1080/25751654.2020.1738815.

Harvey, Adam, and Emile LeBrun. 2023. ‘Computer Vision Detection of Explosive Ordnance: A High-Performance 9N235/9N210 Cluster Submunition Detector’. *The Journal of Conventional Weapons Destruction* 27 (2): 9.

Hasenclever, Andreas, Peter Mayer, and Volker Rittberger. 2000. ‘Integrating Theories of International Regimes’. *Review of International Studies* 26 (1): 3–33. https://doi.org/10.1017/S0260210500000036.

Haykin, Simon S. 2009. *Neural Networks and Learning Machines*. 3. ed. New York Munich: Prentice-Hall.

Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. ‘Unsolved Problems in ML Safety’. arXiv. http://arxiv.org/abs/2109.13916.

Hickey, Samuel M. 2022. ‘Trust but Verify: How to Get There by Using next-Generation Nuclear Verification and Warhead Dismantlement Techniques’. *Bulletin of the Atomic Scientists* 78 (2): 91–97. https://doi.org/10.1080/00963402.2022.2038895.

Holmes, James R., and Andrew C. Winner. 2007. ‘The Proliferation Security Initiative: A Global Prohibition Regime in the Making?’ *Defense & Security Analysis* 23 (3): 281–95. https://doi.org/10.1080/14751790701573881.

Hu, Yupeng, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2023. ‘Artificial Intelligence Security: Threats and Countermeasures’. *ACM Computing Surveys* 55 (1): 1–36. https://doi.org/10.1145/3487890.

Human Rights Watch. 2010a. *Meeting the Challenge: Protecting Civilians through the Convention on Cluster Munitions*. https://www.hrw.org/report/2010/11/22/meeting-challenge/protecting-civilians-through-convention-cluster-munitions.

———. 2010b. *Meeting the Challenge: Protecting Civilians through the Convention on Cluster Munitions*. New York, NY: Human Rights Watch.

———. 2023. ‘Ukraine: Civilian Deaths from Cluster Munitions’. 6 July 2023. https://www.hrw.org/news/2023/07/06/ukraine-civilian-deaths-cluster-munitions.

Hynek, Nik. 2017. ‘Regime Theory as IR Theory: Reflection on Three Waves of’Isms’.’ *Central European Journal of International & Security Studies* 11 (1). https://cejiss.org/images/issue_articles/2016-volume-10-issue-4/11-cejiss-cejiss-0117-electronic.pdf.

———. 2018. ‘Theorizing International Security Regimes: A Power-Analytical Approach’. *International Politics* 55 (3): 352–68. https://doi.org/10.1057/s41311-017-0084-2.

IAEA. 2016. ‘Basics of IAEA Safeguards’. Text. IAEA. 8 June 2016. https://www.iaea.org/topics/basics-of-iaea-safeguards.

———. 2020. ‘Emerging Technologies Workshop: Insights and Actionable Ideas for Key Safeguards Challenges’. https://www.iaea.org/sites/default/files/20/06/emerging-tehnologies-workshop-290120.pdf.

Jervis, Robert. 1982. ‘Security Regimes’. *International Organization* 36 (2): 357–78.

Jervis, Robert, Michal Smetana, and Nik Hynek. 2015. ‘Global Nuclear Disarmament: Strategic, Political, and Regional Perspectives - FOREWORD by Robert Jervis’. In *Global Nuclear Disarmament: Strategic, Political, and Regional Perspectives*.

Kaissis, Georgios, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, et al. 2021. ‘End-to-End Privacy Preserving Deep Learning on Multi-Institutional Medical Imaging’. *Nature Machine Intelligence* 3 (6): 473–84. https://doi.org/10.1038/s42256-021-00337-8.

Keel, Frances M., Steve Lamontagne, Chris A. Pickett, and Keith M. Tolk. 2010. ‘Preliminary Results from the 2010 INMM International Containment and Surveillance Workshop’. In *International Workshop on Containment & Surveillance: Concepts for the 21st Century, Baltimore, Maryland*, 7–11. https://www.ipndv.org/wp-content/uploads/2017/11/Keel_2010_INMM_workshop_on_containment_and_surveillance.pdf.

Keohane, Robert O. 1982. ‘The Demand for International Regimes’. *International Organization* 36 (2): 325–55. https://doi.org/10.1017/S002081830001897X.

———. 1988. ‘International Institutions: Two Approaches’. *International Studies Quarterly* 32 (4): 379–96. https://doi.org/10.2307/2600589.

———. 2005. *After Hegemony: Cooperation and Discord in the World Political Economy*. 1st Princeton classic ed. A Princeton Classic Edition. Princeton, N.J: Princeton University Press.

Keohane, Robert O., and Joseph S. Nye. 2012. *Power and Interdependence*. 4. ed. Longman Classics in Political Science. Glenview, IL: Pearson.

King’s College London. 2016. ‘Research Opens a Window into Pakistan’s Nuclear Weapons Programme’. King’s College London. 4 November 2016. https://www.kcl.ac.uk/news/spotlight/research-opens-a-window-into-pakistans-nuclear-weapons-programme.

Klaise, Janis, Arnaud Van Looveren, Clive Cox, Giovanni Vacanti, and Alexandru Coca. 2020. ‘Monitoring and Explainability of Models in Production’. arXiv. http://arxiv.org/abs/2007.06299.

Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. 2021. ‘Stronger Data Poisoning Attacks Break Data Sanitization Defenses’. arXiv. http://arxiv.org/abs/1811.00741.

Krasner, Stephen D. 1982. ‘Structural Causes and Regime Consequences: Regimes as Intervening Variables’. *International Organization* 36 (2,): 185–205.

———. 1983. *International Regimes*. Cornell University Press.

Kristensen, Hans M., and Matt Korda. 2023. ‘United States Nuclear Weapons, 2023’. *Bulletin of the Atomic Scientists* 79 (1): 28–52. https://doi.org/10.1080/00963402.2022.2156686.

Kristensen, Hans M., Matt Korda, Eliana Johns, and Mackenzie Knight. 2024. ‘Russian Nuclear Weapons, 2024’. *Bulletin of the Atomic Scientists* 80 (2): 118–45. https://doi.org/10.1080/00963402.2024.2314437.

Kubbig, Bernd W. 2005. ‘America: Escaping the Legacy of the ABM Treaty’. *Contemporary Security Policy* 26 (3): 410–30. https://doi.org/10.1080/13523260500500542.

Lacey, Michael O. 2009. ‘Cluster Munitions: Wonder Weapon or Humanitarian Horror’. *Army Lawyer* 2009 (5): 28–33.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. ‘Deep Learning’. *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Levy, Marc A, Oran R Young, and Michael Zurn. 1995. ‘THE STUDY OF INTERNATIONAL REGIMES’. *European Journal of International Relations* 1 (3): 267–330.

Mahesh, Batta. 2020. ‘Machine Learning Algorithms - A Review’. *International Journal of Science and Research (IJSR)* 9 (1): 381–86. https://doi.org/10.21275/ART20203995.

Maxwell, Joseph Alex. 2013. *Qualitative Research Design: An Interactive Approach*. 3rd edition. Applied Social Research Methods Series 41. Los Angeles London New Delhi: Sage.

Meyer, Paul. 2017. ‘The Nuclear Nonproliferation Treaty: Fin de Régime?</Em>’. *Arms Control Today* 47 (3): 16–22.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill Series in Computer Science. New York: McGraw-Hill.

Mullens, James A, Paul A Hausladen, Philip Bingham, Daniel E Archer, Brandon Grogan, and John T Mihalczo. n.d. ‘Use of Imaging for Nuclear Material Control and Accountability’.

Nadelmann, Ethan A. 1990. ‘Global Prohibition Regimes: The Evolution of Norms in International Society’. *International Organization* 44 (4): 479–526.

Paleyes, Andrei, Raoul-Gabriel Urma, and Neil D. Lawrence. 2022. ‘Challenges in Deploying Machine Learning: A Survey of Case Studies’. *ACM Comput. Surv.* 55 (6): 114:1-114:29. https://doi.org/10.1145/3533378.

Papernot, Nicolas. 2018. ‘A Marauder’s Map of Security and Privacy in Machine Learning’. arXiv. http://arxiv.org/abs/1811.01134.

Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow. 2016. ‘Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples’. arXiv. http://arxiv.org/abs/1605.07277.

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. 2018. ‘SoK: Security and Privacy in Machine Learning’. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399–414. London: IEEE. https://doi.org/10.1109/EuroSP.2018.00035.

Patton, Tamara, Sébastien Philippe, and Zia Mian. 2019. ‘Fit for Purpose: An Evolutionary Strategy for the Implementation and Verification of the Treaty on the Prohibition of Nuclear Weapons’. *Journal for Peace and Nuclear Disarmament* 2 (2): 387–409. https://doi.org/10.1080/25751654.2019.1666699.

Price, Richard. 1995. ‘A Genealogy of the Chemical Weapons Taboo’. *International Organization* 49 (1): 73–103.

Quiñonero-Candela, Joaquin, ed. 2009. *Dataset Shift in Machine Learning*. Neural Information Processing Series. Cambridge, Mass.: MIT Press.

Rappert, Brian, and Richard Moyes. 2009. ‘The Prohibition of Cluster Munitions: Setting International Precedents for Defining Inhumanity’. *The Nonproliferation Review* 16 (2): 237–56. https://doi.org/10.1080/10736700902969687.

Revill, James, and María Garzón Maceda. 2022. ‘The Role of Open Source Data and Methods in Verifying Compliance with Weapons of Mass Destruction

Agreements'. In *Open Source Investigations in the Age of Google*, Volume 4:241–58. Security Science and Technology, Volume 4. WORLD SCIENTIFIC (EUROPE). https://doi.org/10.1142/9781800614079_0013.

Ritchie, Nick, and Ambassador Alexander Kmentt. 2021. 'Universalising the TPNW: Challenges and Opportunities'. *Journal for Peace and Nuclear Disarmament* 4 (1): 70–93. https://doi.org/10.1080/25751654.2021.1935673.

Rogers, Jessica, Matt Korda, and Hans M. Kristensen. 2022. 'The Long View: Strategic Arms Control after the New START Treaty'. *Bulletin of the Atomic Scientists* 78 (6): 347–68. https://doi.org/10.1080/00963402.2022.2133287.

Rosenberg, Ihai, Asaf Shabtai, Yuval Elovici, and Lior Rokach. 2021. 'Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain'. arXiv. https://doi.org/10.48550/arXiv.2007.02407.

Rosert, Elvira, and Frank Sauer. 2021. 'How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies'. *Contemporary Security Policy* 42 (1): 4–29. https://doi.org/10.1080/13523260.2020.1771508.

Ruzicka, Jan. 2018. 'Behind the Veil of Good Intentions: Power Analysis of the Nuclear Non-Proliferation Regime'. *International Politics* 55 (3): 369–85. https://doi.org/10.1057/s41311-017-0086-0.

Sauer, Frank. 2016. *Atomic Anxiety: Deterrence, Taboo and the Non-Use of U.S. Nuclear Weapons*. Basingstoke: Palgrave Macmillan. http://link.springer.com/10.1057/9781137533746.

Sauer, Tom, and Mathias Reveraert. 2018. 'The Potential Stigmatizing Effect of the Treaty on the Prohibition of Nuclear Weapons'. *The Nonproliferation Review* 25 (5–6): 437–55. https://doi.org/10.1080/10736700.2018.1548097.

Schelling, Thomas C. 2007. 'The Nuclear Taboo'. *MIT International Review*.

Schenck, Lisa M., and Robert A. Youmans. 2011. 'From Start to Finish: A Historical Review of Nuclear Arms Controls Treaties and Starting over with the New Start'. *Cardozo Journal of International and Comparative Law* 20 (2): 399–436.

Schmidhuber, Juergen. 2015. 'Deep Learning in Neural Networks: An Overview'. *Neural Networks* 61 (January):85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. 'Hidden Technical Debt in Machine Learning Systems'. In *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.

Siracusa, Joseph M., and Aiden Warren. 2018. 'The Nuclear Non-Proliferation Regime: An Historical Perspective'. *Diplomacy & Statecraft* 29 (1): 3–28. https://doi.org/10.1080/09592296.2017.1420495.

Solaiman, Irene, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, et al. 2024. 'Evaluating the Social Impact of Generative AI Systems in Systems and Society'. arXiv. http://arxiv.org/abs/2306.05949.

Tannenwald, Nina. 2005. 'Stigmatizing the Bomb: Origins of the Nuclear Taboo'. *International Security* 29 (4): 5–49.

———. 2007. *The Nuclear Taboo: The United States and the Non-Use of Nuclear Weapons since 1945*. Cambridge Studies in International Relations 87. Cambridge: Cambridge university press.

———. 2012. ‘The Nuclear Taboo’. In *Handbook of Nuclear Proliferation*, edited by Harsh V. Pant, 1st ed., 62–74. Routledge. https://doi.org/10.4324/9780203840849-6.

The Nuclear Threat Initiative. 2014. ‘Innovating Verification: New Tools & New Actors to Reduce Nuclear Risks’, July. https://www.nti.org/wp-content/uploads/2014/07/VPP_Overview_FINAL.pdf.

Thomson, David B. 1999. ‘A Guide to the Nuclear Arms Control Treaties’. Citeseer. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=19a2abfd682ca28a7a7f45512d579c59b90bddfa.

UN General Assembly. 1946. ‘VIII. Resolution: Establishment of a Commission to Deal with the Problems Raised by the Discovery of Atomic Energy’, January. https://documents.un.org/doc/resolution/gen/nr0/032/52/pdf/nr003252.pdf?token=ObyF429oMn6CQzxEkd&fe=true.

Underdal, Arild. 1995. ‘The Study of International Regimes’. Edited by Manfred Efinger, Volker Rittberger, and Michael Zürn. *Journal of Peace Research* 32 (1): 113–19.

United Nations. 1975. ‘Treaty on the Non-Proliferation of Nuclear Weapons’. *Office of Public Information*. https://treaties.unoda.org/t/npt.

———. 2008. ‘Convention on Cluster Munitions’, November.

———. 2021. ‘Treaty on the Prohibition of Nuclear Weapons’, January. https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVI-9&chapter=26.

U.S. Department of State. n.d. ‘Strategic Arms Limitation Talks (SALT II)’. U.S. Department of State Archive. Accessed 19 July 2024. //2009-2017.state.gov/t/isn/5195.htm.

Wan, Wilfred, and Vladislav Chernavskikh. 2023. *Expanding Perspectives on Nuclear Disarmament*. Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-519635.

Wendt, Alexander. 1992. ‘Anarchy Is What States Make of It: The Social Construction of Power Politics’. *International Organization* 46 (2): 391–425.

———. 1994. ‘Collective Identity Formation and the International State’. *American Political Science Review* 88 (June). https://doi.org/10.2307/2944711.

———. 1995. ‘Constructing International Politics’. *International Security* 20 (1): 71–81. https://doi.org/10.2307/2539217.

Williams, Paul D., ed. 2008. *Security Studies: An Introduction*. London ; New York: Routledge.

Wu, Tianyu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. ‘A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development’. *IEEE/CAA Journal of Automatica Sinica* 10 (5): 1122–36. https://doi.org/10.1109/JAS.2023.123618.

Zhang, Angela, Lei Xing, James Zou, and Joseph C. Wu. 2022. ‘Shifting Machine Learning for Healthcare from Development to Deployment and from Models to Data’. *Nature Biomedical Engineering* 6 (12): 1330–45. https://doi.org/10.1038/s41551-022-00898-y.

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. ‘Universal and Transferable Adversarial Attacks on Aligned Language Models’. arXiv. http://arxiv.org/abs/2307.15043.