

Release the bots of war: social media and Artificial Intelligence as international cyber-attack¹

Jasper Schellekens, *University of Malta (Msida, Malta)*

E-mail: jschellekens@gmail.com

ORCID ID: 0000-0002-7351-0090

Abstract

The possibility of conducting attacks on critical infrastructure of states prompted a re-evaluation of the *jus ad bellum* in cyberspace and the drafting of the Tallinn Manual at the behest of the *NATO Cooperative Cyber Defence Centre of Excellence*. Artificial intelligence combined with the use of social media platforms that have access to large audience has opened a new avenue of international dynamics, posing a potential threat to the political independence of states. This article presents the analogy in the use of algorithmic targeting misinformation and influence campaigns and cyber-attacks, as well as examines the roles of the various actors in the international sphere with a view on understanding what actions, if any, nations can undertake to counter these threats to their political independence under international law.

Keywords: information warfare, use of force, artificial intelligence, international law, social media, algorithmic targeting misinformation, influence campaigns, cyber-attacks.

Wojna botów: media społecznościowe i sztuczna inteligencja – międzynarodowy atak cybernetyczny

Streszczenie

Możliwość przeprowadzania ataków na infrastrukturę krytyczną państw skłoniła do ponownej oceny *ius ad bellum* w cyberprzestrzeni oraz opracowania podręcznika (ang. *Tallinn Manual*) na polecenie *Centrum Doskonalenia Cyberobrony NATO*. Sztuczna inteligencja w połączeniu z wykorzystaniem platform mediów społecznościowych, które mają dostęp do dużej liczby odbiorców, otworzyły nową drogę dynamiki międzynarodowej, stwarzając potencjalne zagrożenie dla niezależności politycznej państw. Niniejszy artykuł dokonuje analogii w stosowaniu algorytmicznego kierowania kampaniami dezinformacyjnymi i influencerskimi oraz ataków cybernetycznych. Analizuje role różnych aktorów

¹ The opinion expressed in this article are the author's own opinion, and it is not represent in any manner this journal or the University of Warsaw. The conclusions and assessments in this article are based on the knowledge and experience of the author.

w sferze międzynarodowej w celu zrozumienia, do czego, jeśli w ogóle, mogą zobowiązać się narody w imię przeciwdziałania tym zagrożeniom dla ich niezależności politycznej na mocy prawa międzynarodowego.

Słowa kluczowe: wojna informacyjna, użycie siły, sztuczna inteligencja, prawo międzynarodowe, media społecznościowe, algorytmiczne kierowanie kampaniami dezinformacyjnymi, kampanie influencerskie, ataki cybernetyczne.

Artificial intelligence (AI) is an important addition to the tactical arsenal of the military operations, with big data assisting in field decisions and logistics, and the increase deployment of AI tools in weapons (Mizokami 2020). It stands to be a pillar of military operations in the near future for the United States and China looking to harness the advantages AI could provide the military (U.S. strategy, China's strategy) and significant research directed at the development of autonomous weapons. While AI can be wielded like a weapon, it is perhaps best viewed enabler of military force, in the same way as more general technological advances were such as electricity, radios, and radar (Johnson 2019: p. 150). And in a similar vein as these technologies it seems likely to revolutionize the way people conduct war. Some scholars, such as Kenneth Payne and Thomas Elkjer Nissen, go so far as to say that these will result in new cognitive processes and new ways of conducting warfare (Payne 2018; Noël 2018; Nissen 2015). China and Russia who have traditionally been more cognisant of the impact of information (Stinissen 2015; Kania 2019) see the development of military AI technologies as an arms race, the winner of which will have an important advantage in the global military arena.

Combining Artificial Intelligence with social media

While there is a growing scholars' interest on the use of AI for military purposes, a lot of research are focused on the development of autonomous weapons or logistics. Nevertheless, the key feature enabled by AI, namely the rapid diffusion and synthesis of information, has received less academic attention (Johnson 2019: p. 148) and is critical feature of information warfare. Information warfare is a fairly wide range of topic and the scholars have approached it from a number of areas. In this article, the use of AI and social media will be the central element of information warfare examined. Information warfare is not a new concept (Prier 2017), and the media and the social world have always combined to contribute to a mediated construction of the social reality, but it merits investigation due to "the pervasive influence of social media on daily lives; and the visible changes underway within practices and infrastructures of mainstream media outlets" (Grech 2019: p. 215).²

The key international treaties, such as the UN Charter, where article 4 is an important legal provision on the use of force and the Geneva conventions, were last revised in the

² For details of what can be included in information warfare see the publication by Ch. Bellamy, who wrote: "These include information-age warfare, information-based warfare, electronic warfare and command and control (C2) warfare, a more manageable shorthand for C4I (command, control, communications, computers, and intelligence)." (Bellamy 2001: p. 56)

wake of the Second World War. The discussion about the use of force in cyberspace is not a new one, with the *Tallinn Manual* (see: Schmitt 2013) and *Tallinn Manual 2.0* (see: Schmitt 2017) emerging as a sort of guide as to how the experts viewed the status of the use of force in cyberspace. However, since the publication of the *Tallinn Manual* the use of artificial intelligence in media has grown alongside the size of data available and the growth of social networking platforms such as Facebook and Twitter. In the time since the *Tallinn Manual* was published in 2013, Facebook has grown from 1,228 million of active users in Q4-2013 to 2,797 million at the end of 2020 (Statista 2021), while not as dramatic as the increase of Twitter's users – 100 million between 2013 and 2020³ (see details: Statista 2019). There is also further recognition by scholars that artificial intelligence can pose a threat to International Humanitarian law, even when not directly linked to weapon systems (Klonowska 2020). Thomas Elkjer Nissen has further highlighted the complications arising from social media use for military purposes as it challenges the normative system of wartime, where International Humanitarian Law (IHL) and customary international law of war apply, and peacetime, where domestic law and civil protection apply (Nissen 2015: p. 114). Using social media to create specific effects, such as mobilising civilians, requires sophisticated planning (Nissen 2015: p. 58), with ever-increasing amounts of data and ever increasing targeting possibilities, AI makes this planning easier both by developing automation process (e.g. botnets) as well as taking advantage of the existing automation processes (e.g. hashtag poisoning) to gain military advantages. The combination of social media and AI – primarily through bots and algorithmic manipulation – has already had a significant impact on the 2016 U.S. presidential elections (Allcott, Gentzkow 2017; Bakir, McStay 2017; Vargo et al. 2018; Guess et al. 2020), the European elections in Italy (Pierrri et al. 2020), the German federal elections (Morstatter et al. 2018), the British and Dutch elections in 2010 (Broersma, Graham 2012), and militarily specifically during the Russian annexation of Crimea in 2014 (Jaitner 2015; Lange-Ionatamishvili, Svetoka 2015; Blank 2017; Stinissen 2015).

An unexpected line of attack?

Examining the military use of a combination of AI and social media is to a large extent analogous with cyber-warfare (Nissen 2015: p. 115), that is why I will be looking to extend the concepts and rules set out in the *Tallinn Manual 2.0* on cyber warfare and explore to what degree these can be applied to "weaponized social media" in the field of warfare and armed conflicts as well as to explore the kind of countermeasures the existing framework of international law and IHL allow or facilitate in the case of weaponized social media. The intent is not a detailed application of *jus ad bellum* or *jus in bello* to attacks conducted via Social Media and with the use of AI, because such inquiries are fact specific. Instead, it attempts to use the analogous thinking behind cyber warfare as

³ Twitter went from 241 million users in 2013 to 353 million users at the end of 2020. After this they switched measuring only "monetizable users". It is also important to note that the vast majority of their users are US-based.

set out in the *Tallinn Manual* as a starting point to determine the legality of some of the actions and the possibilities of states to react on the basis thereof.

Firstly, I will show that the use of AI and social media can indeed be considered as an "attack". This will be done by drawing on the *Tallinn Manual*, as well as customary international law. In the first part, the important related legal concepts will be foregrounded (such as sovereignty, use of force, and countermeasures).

Secondly, I will examine how grievous attacks through social media can be considered. In particular, whether attacks via use of AI and social media could be a violation of international law and whether it can ever reach the threshold to be considered as the use of force under international law.

These will be illustrated with case studies and examples, primarily focusing on the most clear-cut case of the annexation of Crimea. This is one of the few territorial invasions that has occurred in Europe since the rise of social media, and it shows a case study of what weaponized social media entails. These parts do not attempt to offer a conclusion about where the lines ultimately should be drawn, but it discusses merits and challenges of interpretation.

Thirdly, I will highlight some of the most salient points and offer some guidance to the interpretation, leading into the conclusion which suggests avenues for expansion and further research.

War in cyberspace – applying the rules of the *Tallinn Manual* to AI and social media

The *Tallinn Manual* is an interpretation by international law experts, taking into account the relevant treaties in addition to customary international law, but also emerging state practice, which makes it an ideal starting point for analysis. Technology develops and states' positions shift as does the relevant international law, hence the CCDCOE has started a five-year project to update to the *Tallinn Manual 3.0* (see: CCDCOE 2021). As it takes into account an expert consensus on the topics at hand it provides for quick reference guide for the legal principles associated with cyber warfare. In the following section, I have divided the *Tallinn Manual* rules that are applicable in cyber operations and extended them to the combination of social media and AI. I've left out espionage, as through this could potentially contribute to gaining local access and potentially influencing local cyber-physical systems. Espionage in specific would fall under the domestic law of the country and therefore wouldn't fall under international law governing use of force.

Attacks and cyber-attacks

The *Tallinn Manual* defines a cyber-attack as "a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects" (Schmitt 2013: p. 106). Elaborating on the definition the manual distinguishes that it is the "use of violence" that qualifies an operation as an "attack" and it

explicitly states that "non-violent operations, such as psychological cyber operations or cyber espionage, do not qualify as attacks" (Schmitt 2013: p. 92).

The majority of the International Group of Experts agreed that the notion of armed attack did not necessarily involve the employment of weapons, but instead focused on the effects of the attack (Schmitt 2013: p. 54). Their position is further strengthened by Boothby, who argues: "a weapon is an offensive capability that is applied, or that is intended or designed to be applied, to a military object or enemy combatant. A destructive, damaging or injurious effect of the weapon need not result from physical impact as the offensive capability need not be kinetic" (Boothby 2014: p. 175). "With this definition Boothby also implies, although in connection with a broader debate on "cyber-weapons", that social network media, in a "weaponized" form, also can be regarded as weapons and therefore need to be assessed as such under the rules and provisions of international humanitarian law" (Nissen 2015: p. 116; see also: Klonowska 2020).

The International Group of Experts was divided on their opinion about where the line was drawn to label of "armed attack". For example, the fact that "the Syrian Electronic Army hacked the Twitter account of the Associated Press and posted a false rumor about a terror attack on the White House" (Ferrara et al. 2016: p. 99), would be interpreted as an armed attack (Schmitt 2013: p. 55).

From the perspective of customary international law, it is telling that no international cyber incidents have "unambiguously" and "publicly" been characterised as reaching the threshold of an armed attack by states, not even when referencing the "cyber war" between Russia and Estonia in 2007 (Schmitt 2013: p. 43-44, 56).

Hathaway et al. (2012: p. 821) propose a broader definition of cyber-attacks describing "three common forms: distributed denial of service attacks, planting inaccurate information, and infiltration of a secure computer network". The reference to planting of inaccurate information as a form of cyber-attack is particularly relevant to operations undertaken through the use of social media. The *Tallinn Manual* rules indicate that unless an injury or death occurs or there is damage or destruction of objects, the operation cannot be classified as a cyber-attack. However, real harm to national security can occur with little or no direct physical consequences, in line with many effects-based legal scholars such as Hathaway et al. (2012), Nissen (2015), and Schmitt (2013, 2014), the effect of military use of social media may result in harm. However, the type of harm is more difficult to quantify (e.g. effect on democratic process, effect on political independence, national confusion, etc.) and the causal link more difficult to establish.

Sovereignty

Respect for sovereignty is one of the pillars of the UN Charter and customary international law, Article 2(4) of the UN Charter provides protection from the "threat or use of force against the territorial integrity or political independence of any State".

Sovereignty impacts the use of social media and AI for military purposes as a State's political independence can be attacked using a combination of social media and AI

(Allcott, Gentzkow 2017; Bakir, McStay 2017; Vargo et al. 2018; Guess et al. 2020; Pierrri et al. 2020; Broersma, Graham 2012). However, sovereignty also obliges a state to not knowingly allow its territory to be used for acts contrary to the rights of other states. The nearly endless ways for information to flow results in a challenge for states to prevent their territories from being used to route data and packets that could eventually lead to a cyber-attack and the exact extent of this obligation is still undetermined (Shackelford et al. 2016: p. 21). Data and packets that travel through social media channels are even more elusive as these do not even target state-owned infrastructure (e.g. power plants, smart grids, automated transportation infrastructure). In this article, the analysis is focused on the attacks on sovereignty experience by the state, without in-depth consideration of the obligations of the transit states, because attacks making use of the social media will be virtually indistinguishable from other commercial and personal use of the network.

The *Tallinn Manual* outlines that "the principle of sovereignty allows a State to, *inter alia*, restrict or protect (in part or in whole) access to the internet, without prejudice to applicable international law, such as human rights or international telecommunications law"⁴ (Schmitt 2013: p. 26). However, this exercise of sovereignty results in a self-denial, which may cause further harm to both states, either financial or political (Schmitt 2013: p. 33).

The majority of the International Group of Experts also agreed on an effects-based approach, meaning that if the effects were analogous to those of a kinetic armed attack a cyber operation could be qualified as an attack (Schmitt 2013: p. 54). However, the harm resulting from political interference or an attack on the sovereignty of the state itself is difficult to quantify but can potentially have more devastating effect than an attack on critical infrastructure. Democratic systems rely heavily on information (Whyte 2020b) and the key elements of democratic information are the origin of information, credibility of course, quality of information and freedom of information (Whyte 2020a). Social media is proven to have an impact on voter behaviour as evidenced by the study published in *Nature magazine* that calculated approximately "340,000 extra people turned out to vote in the 2010 US congressional elections because of a single election day Facebook message" (Corbyn 2012; see also: Bond et al. 2012: p. 297). As Broniatowski et al. (2018: p. 1381) point out in their study on weaponized health communication, Twitter bots and trolls were demonstrated to be effective linking health communication to "divisive topics in US culture", thereby using the built-in functionality of algorithm for rapid dissemination as well as serving to polarise the internal national discourse. The contribution of bots to the outcome of Brexit and the 2016 U.S. Presidential Election is may seem negligible, but as Whyte points out, this is enough for elections that are contested on such small margins (Whyte 2020a).

However, a cyber operation may still be unlawful under international law, even when it does not fulfill the criteria to be labeled as an armed attack or as use of force. While there is some evidence that the concept of sovereignty is adapting (Jacobsen et al. 2016; Paris 2020), as it stands enshrined in the principle of the sovereign equality of states laid out in

⁴ For example, the ITU Constitution.

Article 2(1) of the United Nations Charter the prohibition of intervention is implicit (Schmitt 2013: p. 46; International Court of Justice 1986: par. 202). Therefore, a state is entitled to undertake countermeasures under international law (Schmitt 2014; Shackelford et al. 2016).

Use of force and countermeasures

The threshold for a cyber operation to be considered an armed attack equivalent to the use of force is high, with Rule 11 of the *Tallinn Manual* qualifying that it needs to be "comparable to non-cyber operations rising to the level of a use of force" (Schmitt 2013: p. 47). Even in the cases where states were involved in "cyber warfare", none of the states has unambiguously labeled the cyber operations as "armed attacks" or as a "use of force". This aligns with both the *Tallinn Manual* and Schmitt's own arguments against determining that cyber-attacks – and by analogy social media attacks – fulfill the requirements to be considered as use of force.

Tallinn Manual underlines factors that states are likely to consider when judging whether a cyber operation constitutes a use of force such as "severity, immediacy, directness, invasiveness" (Schmitt 2013: p. 49). Consideration of these factors will always make attacks conducted via social media impossible to respond to through use of force. The links between the information campaign and any subsequent events will always be too tenuous for these factors. This is further reinforced by the Experts in the *Tallinn Manual* making clear in no uncertain terms that "a false tweet (Twitter message) sent out in order to cause panic, falsely indicating that a highly contagious and deadly disease is spreading rapidly throughout the population" does not qualify neither as an attack nor threat thereof (Schmitt 2013: p. 105). Even when evidence can be provided of direct impact, which is rare due to the extensive use of non-state actors, it is even more difficult to attribute the actions to any given state. Including economic or political coercion as "force" was considered but ultimately rejected during the Charter drafting conference in San Francisco. And consistent with this the *Tallinn Manual* also states: "Cyber operations that involve, or are otherwise analogous to, these coercive activities are definitely not prohibited uses of force" (Schmitt 2013: p. 47–48).

Social media attacks and cyber attacks can, and often do, happen in parallel with military activity on the ground, which can give rise to the situation that the military use of social media would fall under the law of armed conflict. In this case it is interesting that while a false Tweet was unequivocally ruled out as an attack, the majority of the International Expert Group did consider it "would be unlawful if the operation undermined the principle of distinction (Rule 31) by placing civilians and civilian objects at increased risk" (Schmitt 2013: p. 153).⁵

A state need not be injured by a use of force to have recourse to countermeasure under international law (Schmitt 2013: p. 26). In order for a state to take countermeasures there must be (1) a breach of international law obligations in respect to the injured state,

⁵ The example mentioned in the *Tallinn Manual* concerns a military computer system using the .com domain in order to appear to be commercial in nature making it harder to detect.

and (2) the wrongful act needs to be attributed to the responsible state (Schmitt 2014; Hathaway et al. 2012: p. 858).

Schmitt refutes the "prevailing sense that States stand defenseless" unless a cyber operation is qualified as an armed attack as dangerous to international peace and security (Schmitt 2014: p. 730). The countermeasures available under international law are sufficient to address cyber attacks an elevating cyber operations to use of force or armed attacks risks that response may take on more forceful forms (Schmitt 2014: p. 730).

Countermeasures may only be taken by a state against another state, which may result in some challenges both in terms of attribution, as well as how states can react to attacks launched by non-state actors (Schmitt 2014: p. 731). The *Tallinn Manual* allows for protective measures to be executed by a state if faced with cyber incidents and if there is no other way to address the situation, even if the origin of the incident is unclear (Schmitt 2013: p. 43). In 2007 the Estonian CERT responded to the cyber threat by suspending service to IP addressed from Russia, because the ITU Constitution and the notion of sovereignty both allow a state to stop international telecommunications within their own territory this does not qualify as a countermeasure under international law.

Social media warfare in action – an examination of practical aspects

Information has long been used as a weapon by Russia for mobilising its own populations, as well sowing mistrust of foreign powers. In fact, Russian Maj. Gen. (R) Ivan Vorobyov and Col. (R) Valery Kiselyov clearly stated in their publication in military journal *Voyennaya Mysl – Military Thought* that information is a type of weapon (see: Jaitner 2015: p. 87). The Deplorable Network included a bot network of between 16,000 and 34,000 accounts and played a significant role in exerting Russian influence over the 2016 U.S. elections (Prier 2017: p. 73).

As discussed above, using social media there are cases that can cause harm, but do not meet the thresholds to be considered an armed attack or use of force. By leveraging the targeting effectiveness of social media and the related algorithms, states can make use of these platforms for military purposes. Use of botnets and third parties further complicate conclusively tracing the origin of these attacks, but evidence strongly points to a coordinated Russian operation (Blank 2017: p. 85).

Estonian authorities suspected that the Russian cyber-attacks of 2007 were aimed to incite large enough demonstrations to provoke violence (Blank 2017: p. 86). Demonstrations of the desired magnitude never manifested and Russia was disappointed by the lack of support of Russians living in Estonia (Blank 2017: p. 87).

During the annexation of Crimea in 2014, the decisive conflict was in the cyber and communications domain, and the result of the conflict was determined "without firing a single shot" (Lange-Ionatamishvili, Svetoka 2015: p. 103). Again, as was the case years earlier in Estonia, Russian bots and operatives, used social media to mobilise the local population in Ukraine (Blank 2017: p. 86). Preparations started much earlier to ensure mobilisation with the registration of the official websites for the People's Republics of

Donetsk and Lugansk, and "Novorossiya websites such as novorus.info and novorossia.su" (Jaitner 2015; p. 92).

While a direct link is difficult to ascertain, it has been shown by the influence of bots on the vaccine discourse and the 340,000 additional voters in the U.S. elections that social media is capable of mobilising people, if not changing their minds. This influence as we've seen reflected in political advertising is referred to as "nudging" and can be harnessed to alter behaviour and is mainly about facilitating the choices (Nissen 2015; p. 85). Combined with extremely accurate targeting algorithms, it is plausible, if not likely that social media is able to mobilise protests or other political action. The targeted marketing further allows for techniques and technologies that exploit cognitive limitations of their targets (Nadler et al. 2018; Calo 2014). Therefore, it is no surprise that the conflict in Eastern Ukraine started from protests at Maidan Square and led to the occupation of Crimea (Stinissen 2015; p. 131).

The targeting function of AI and social media allows for states to target specific subsections of another state, often with the ability to profile along their political beliefs, for example "anti-Semitism" (Angwin et al. 2017). In 2021 Facebook, recently re-branded as Meta, has removed a number of advertising target options that are considered "sensitive" (Silberling 2021). In the case of Crimea, a referendum was held, "which declared 97% of the voters supported joining Russia" (Stinissen 2015; p. 126). In principle, as we have seen the capabilities of "nudging" and mobilisation *via* social media, it could in theory permit states to target minority groups, aggravating national conflicts, and in the case of Russia and Crimea specifically, exploit the fundamental right to self-determination in order to acquire territory legally in possession of another state (Burke-White 2014).

Russian president V. Putin defended the independence of Crimea invoking the legal principle of self-determination and referring to a decision of the International Court of Justice and a UN Security Council Resolution general international law contains no prohibition on declarations of independence. Any direct link of influence on the mobilisation and activities of the Donetsk and Lugansk separatists is tenuous and the threshold set out for support to be labeled as a use of force is high as International Court of Justice held that supplying funds to insurgents was "undoubtedly an act of intervention in the internal affairs of Nicaragua", although not a use of force (see: International Court of Justice 1986: p. 119, par. 228; Schmitt 2013; p. 46). Therefore, even where there is a combination of social media and conflict outside of the cyber realm, is it not enough to consider the actions by separatists and Russian financing, training, equipment and operational support, but Russia must also have a role in organising, coordinating, and planning their operations (Stinissen 2015; p. 130). However, through social media "nudging" and targeting using external organisations, Russia does not even need to be involved in organising, coordinating, and planning their operations in order to significantly impact and influence them. While it may still qualify as an intervention allowing for countermeasures (though not use of force) the degree of control remains a challenge because using social media makes it even more difficult to attribute to a state at all. When the requirements for countermeasures reach critical levels, it is often already too

late, as the necessary influence has already been exerted for months or possibly years, as shown by the earlier registrations of domains and targeted advertisement. Thresholds for countermeasures, such as attributing blame to a state, will take too long to reach, and when the critical level is reached it is often too late to have a significant impact. Stinissen concludes that ultimately "cyber operations in the Ukraine conflict have been used either to gather intelligence or as part of an ongoing 'information war' between the parties" (Stinissen 2015: p. 134). As such, they were not designed to inflict damage or injure persons, therefore in accordance with the *Tallinn Manual's* rules on cyber operations and the use of force, which we are looking to extend to that information warfare they have not risen to the level of armed conflict (Stinissen 2015: p. 134). However, this is problematic as we have seen in Ukraine that the information manipulation and the mobilisation of local target groups formed the precedent for the annexation of Crimea.

In April 2021 the U.S. Department of Treasury has issued sanctions against a number of Russian tech companies. Although the sanctions reference a number of specific cyber-attacks such as the attack on the platform *SolarWinds Orion*, the sanction also mentions specifically "facilitating malicious cyber activities against the United States and its allies and partners that threaten the free flow of information" (U.S. Department of the Treasury 2021). This is an important recognition of the importance of information in democratic societies and the impact that it had in this particular series of conflicts.

This is where the largest threat of social media attacks lies: on the political independence of states, where indirectly it threatens the territorial integrity as it can target separatist movements or other unsatisfied minorities in a state. This is a similar turn of events that preceded the storming of the Capitol in 2019, targeting adverts, using bots, misinformation, and key figures on social media, a disgruntled minority was targeted and 'nudged' into action (see: Kaplan 2021; Walsh 2021; Hunt 2021).

Conclusions and recommendations

Social media and the automation facilitated through artificial intelligence has a very real impact on the conduct of hostilities, but also on the political independence of states. Bots are one of primary tools used to conduct cyber and communication operations amplifying fringe actors while creating an artificial sense of relevance and according to media reports they have been deployed by government actors in Argentina, Azerbaijan, Iran, Mexico, the Philippines, Russia, Saudi Arabia, South Korea, Syria, Turkey, and Venezuela (Bradshaw, Howard 2017: p. 11). Russia has a long history of information "warfare" that extends beyond the cyber domain targeting physical, logical, and social layers of society (Jaitner 2015: p. 91). Therefore, it was well-prepared to use information warfare to achieve their objective of annexing Crimea, to such a degree that it managed to circumvent the prohibition on the use of force.

Russia's dedication to information superiority in warfare is evident from the "time and resources that have been spent in creating official, semi-official, and unofficial sources of war-related information, including dedicated channels on YouTube" (Jaitner 2015: p. 91).

The United States has also increased its investment in information security with multiple programmes valued collectively over 52 million dollars, indicating a renewed focus on the use of social media for public opinion manipulation (Bradshaw, Howard 2017 p. 21). The importance attributed to this is highlighted by the wording of the U.S. Department of Treasury sanctions that specifically references that they are a response to, inter alia, the disruption of free flow of information (U.S. Department of the Treasury 2021).

Responsibility of the individuals

Social media is an important source of information. In the United States, 62% of adults get their news from social media (Loos, Nijenhuis 2020: p. 80). Often interactions with fake news or targeted advertisements are attributed to an individual level of responsibility. Where bots are designed to amplify a specific message or add legitimacy to fake news generated for a social media offensive, humans have the luxury of critical thinking and free will. However, studies have demonstrated that overall individuals are easily duped by information on social media (Loos, Nijenhuis 2020: p. 80). Judging information credibility is possible and countering social media disinformation operation with accurate open source analysis is possible, but not necessarily in a timely manner (Jaitner 2015: p. 93). On an individual basis education is key with numerous educational programmes for AI literacy for children, such as the ArtBot project implemented by the Institute of Digital Games (Voulgari et al. 2021).

In this area the European Union is dedicating research and funds to making individual citizens more resilient to this type of social media manipulation finding that "legal restriction of content may pose a greater harm to democracy than disinformation itself" (Bayer et al. 2019: p. 10). However, even with more AI and social media literacy, targeting takes advantage of "human psychological traits and social engineering" (Bayer et al. 2019: p. 12; see also: Nadler et al. 2018; Calo 2014) so that it may be difficult to counter on an individual level. Hence, in addition to any educational curriculum that offers citizens resilience to these types of information operations, actual countermeasures and international mitigation measures need to be considered.

Responsibility of Artificial Intelligence

There are many ways, how AI can be used in combination with social media to make a significant impact that can lead to harm, although there does not seem to be a condition where this harm can reach the threshold required to be considered a use of force. AI can impact information offensives, primarily through the use of bots and through the use of algorithmic targeting (either through targeting a particular demographic or targeting a particular message). An example of the vulnerability of social media to the hijacking of the algorithm to manipulate messages (Pfeffer et al. 2018; Morstatter et al. 2018), but also simply leveraging human psychological traits can also "nudge" people to act (Nadler et al. 2018; Bayer et al. 2019).

AI's feature is not unique for social media, because this type of AI manipulation also occurs in automated commercial outlets such as the algorithms of Amazon and Netflix.

In their proposal for AI regulation published in April 2021 the European Union also recognises the hazards AI can have when combined with social media, proposing the prohibition of AI systems that contravene Union values or have the "significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific vulnerable groups such as children or persons with disabilities in order to materially distort their behaviour in a manner that is likely to cause them or another person psychological or physical harm" (European Commission 2021a: p. 12-13). More cautiously it indicates that other manipulative practice could be governed by existing legislation, but the proposal does seem to recognise the significance of AI in affecting behaviour. The EU also further proposes in the document *Regulatory framework proposal on artificial intelligence* to regulate AI with a risk-based approach, dividing the risks into the following 4 categories: (1) unacceptable risk, (2) high risk, (3) limited risk, and (4) minimal risk (see: European Commission 2021b). Under the high risk AI systems it includes: critical infrastructures, safety components of products; employment, essential private and public services; law enforcement that may interfere with people's fundamental rights, migration, asylum and border control management, and administration of justice and democratic processes (European Commission 2021b). It is unclear under which category social media would fall, as in the current context of communication it might fall under critical information infrastructure, but this remains to be seen.

Regulation of social media

Helberger cautions that regulation and formalising role of social media may actually increase their political power (Helberger 2020: p. 848). This is also in line with the opinion by the experts of the study on disinformation commissioned by the European Parliament (see: Bayer et al. 2019). At the time of writing Helberger notes that there were no proposals to limit the use of AI in social media to persuade for own political purposes (Helberger 2020: p. 849), however since then Facebook has agreed to reign in its more "sensitive" categories (Silberling 2021). These self-imposed restrictions may be a way for the company to avoid stronger regulation and scrutiny rather than a genuine attempt at regulating the political power of the platform. It is of course hard to believe a corporation would willingly put aside the profit resulting from this type of information operations (Nadler et al. 2018).

There is, however, a growing recognition in legal scholarship and precedent that social media mobilisation can in fact have real and direct consequences. For instance the case where "an editor and two journalists were convicted by the International Criminal Tribunal for Rwanda (ICTR) for incitement to commit genocide through their radio broadcasts" (Nissen 2015: p. 117). This can be extended to algorithmically mediated content on social network media as Nissen suggests, although the challenge would remain in tracing down the exact origin of the content. The careful selection of words by the U.S. Department of the Treasury for the imposition of sanctions on Russia and Russian companies, also indicates some recognition of the significance of an information offensive via social media.

In development

Each aspect raised in this overview deserves a more in-depth exploration and there are important aspects related to *jus ad bellum* and *jus in bello* that have not been addressed at all, such as the qualification that social media functions as a command center for certain actors, using information retrieved from the platform to mobilise civilians (entering into discussions of the blurred distinctions between civilians and combatants) or even to target enemy combatants (entering into discussion to the minimum requirement of military targeting systems under international humanitarian law).

The norms of international law in respect to this new 'theatre' of digital conflict are still developing, as Reisman highlights: "International law is still largely a decentralized process, in which much lawmaking (particularly for the most innovative matters) is initiated by unilateral claim, whether explicit or behavioral. Claims to change inherited security arrangements [...] ignite a process of counterclaims, responses, replies, and rejoinders until stable expectations of right behavior emerge." (Reisman 2003: p. 82).

What is of particular interest is that no states engaged in conflicts, where the battle played out on social media have referred to this as a form of cyber-attack. This is also in line with the conclusions of the analysis, indicating that there does not appear to be any situation, where operations conducted via social media would rise to a high enough threshold to be considered a use of force. In the line of reasoning set out by Waxman it may indeed be beneficial for stability that conflicts have shifted from the physical battlefield to a conflict of ideas (Waxman 2011). The problem arises when it doesn't remain solely a conflict of ideas, such as during the annexation of Crimea.

In modern interstate conflict, social media facilitates the targeting of particular groups and the threshold for control established in the Nicaragua judgment by the International Court of Justice in 1986 is high. Therefore, foreign governments have the opportunity to more easily target separatist movements without necessarily being involved directly in any form. Judges Schwebel and Jennings dissented from the threshold established in the judgment, with Jennings stating in their dissenting opinion that the Court's decision dangerously restricts self-defense against support for rebels (International Court of Justice 1986), something that was evident from the operations in Crimea.

The current *status quo*, does allow states to adopt limited, internationally permissible countermeasures that do not rise to the level of armed attacks, such as the sanctions imposed by the United States on Russia following a number of cyber incidents (U.S. Department of the Treasury 2021), which has proven sufficient deterrence thus far (Schmitt 2014). International efforts to mitigate the dangers of social media operations begin with an agreement on the problem and the challenge of harmful attacks, which to not cross the threshold needed for action cannot be met by domestic reforms alone.

Education is an important pillar in order to build a society that is more resilient to the both the algorithmic targeting as well as the psychological techniques these actors are using to mobilise people. But even after being educated on misinformation, people still struggle to discern truth from fiction (see: Loos, Nijenhuis 2020). With the development of Big Data and artificial intelligence, more and more is possible in terms

of tailoring content to a specific audience to mobilise them and in this context *Tallinn Manual* update is timely.

Jasper Schellekens – Research Support Officer at the Institute of Digital Games, University of Malta. He has an LLM from University of Leiden, and his early research were focused on the legality of antisatellite weapons. He has worked for the *Commonwealth Cybercrime Initiative*, helping Commonwealth countries improve their capacity to combat cybercrime starting from the legal frameworks.

Jasper Schellekens – jest pracownikiem ds. wsparcia badań w Instytucie Gier Cyfrowych na Uniwersytecie Maltańskim. Posiada LLM z Uniwersytetu w Leiden, a jego wcześniejsze badania koncentrowały się na legalności broni antysatelitarnej. Pracował dla *Commonwealth Cybercrime Initiative*, pomagając krajom *Commonwealth'u* we wzmocnieniu ich zdolności do zwalczania cyberprzestępczości, począwszy od ram prawnych.

Acknowledgements: The author would like to thank Dr Krista Bonello Rutter Giappone and Beatrice Jacuch for their feedback on early drafts.

➔ References:

- ALLCOTT Hunt, GENTZKOW Matthew (2017), *Social Media and Fake News in the 2016 Election*, "Journal of Economic Perspectives", vol. 31, no. 2. DOI: 10.1257/jep.31.2.211
- ANGWIN Julia, VARNER Madeleine, TOBIN Ariana (2017), *Facebook Enabled Advertisers to Reach 'Jew Haters'*, <https://www.propublica.org/article/facebook-enabled-advertisers-to-reachjew-haters> (14.09.2017)
- BAKIR Vian, MCSTAY Andrew (2017), *Fake News and The Economy of Emotions: Problems, causes, solutions*, "Digital Journalism", vol. 6, p. 154-175. DOI: 10.1080/21670811.2017.1345645
- BAYER Judit et al. (2019), *Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and its Member States*, European Parliament, PE 608.864 – February 2019. DOI: 10.2139/ssrn.3409279
- BELLAMY Christopher (2001), *What is Information Warfare?*, in: Ron Matthews, John Treddenick (eds), *Managing the Revolution in Military Affairs*, London. DOI: 10.1057/9780230294189_4
- BLANK Stephen (2017), *Cyber War and Information War à la Russe*, in: George Perkovich and Ariel E. Levite (eds), *Understanding Cyber Conflict: Fourteen Analogies*, Washington
- BOND Robert M. et al. (2012), *A 61-million-person experiment in social influence and political mobilization*, "Nature", vol. 489. DOI: 10.1038/nature11421
- BOOTHBY William H. (2014), *Conflict Law*, T.M.C. Asser Press, The Hague. DOI: 10.1007/978-94-6265-002-2
- BRADSHAW Samantha, HOWARD Philip N. (2017), *Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation*, Computational Propaganda Research Project, Working paper no. 2017.12, <http://governance40.com/wp-content/uploads/2018/11/Troops-Trolls-and-Troublemakers.pdf> (31.12.2017)
- BROERSMA Marcel, GRAHAM Todd (2012), *Social Media as Beat: Tweets as a news source during the 2010 British and Dutch elections*, "Journalism Practice", vol. 6, p. 403-419. DOI: 10.1080/17512786.2012.663626

- BRONIATOWSKI David A. et al. (2018), *Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate*, "American Journal of Public Health", vol. 108. DOI: 10.2105/AJPH.2018.304567
- BURKE-WHITE William W. (2014), *Crimea and the International Legal Order*, "Survival", vol. 56, issue 4, p. 65–80. DOI: 10.1080/00396338.2014.941548
- CALO Ryan (2014), *Digital Market Manipulation*, "The George Washington Law Review", vol. 82, no. 4.
- CCDCOE (2021), *The Tallinn Manual @ONLINE*, NATO Cooperative Cyber Defence Centre of Excellence. <https://ccdcoe.org/research/tallinn-manual/> (29.11.2021)
- CORBYN Zoe (2012), *Facebook Experiment Found to Boost U.S. Voter Turnout*, "Nature", 12.09.2012. DOI: 10.1038/nature.2012.11401
- EUROPEAN COMMISSION (2021a), *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Proposal for a Regulation of the European Parliament and of the Council*. COM(2021) 206 final, 21.04.2021, Brussels.
- EUROPEAN COMMISSION (2021b), *Regulatory framework proposal on artificial intelligence*, <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (07.12.2021)
- FERRARA Emilio et al. (2016), *The Rise of Social Bots*, "Communications of the ACM", vol. 59, no. 7, p. 96–104. DOI: 10.1145/2818717.
- GRECH Alex (2019), *The Unbearable Lightness of Online Social Networks and the Hyperlocal*, in: J. Borg, M. Lauri (eds), *Navigating the Maltese Mediascape*, Kite Publishing
- GUESS Andrew M., NYHAN Brendan, REIFLER Jason (2020), *Exposure to untrustworthy websites in the 2016 US election*, "Nature Human Behaviour", issue 4, p. 472–480. DOI: 10.1038/s41562-020-0833-x
- HATHAWAY Oona A., CROTOF Rebecca, LEVITZ Philip, NIX Haley, NOWLAN Aileen, PERDUE William, SPIEGEL Julia (2012), *The Law of Cyber-Attack*, "California Law Review", vol. 100 (4), p. 817–886.
- HELBERGER Natali (2020), *The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power*, "Digital Journalism", vol. 8, issue 6, p. 842–854. DOI: 10.1080/21670811.2020.1773888
- HUNT Jennifer S. (2021), *Countering cyber-enabled disinformation: implications for national security*, "Australian Journal of Defence and Strategic Studies", vol. 3, no. 1. DOI: 10.51174/AJDSS.0301
- INTERNATIONAL COURT OF JUSTICE (1986), *Case Concerning Military and Paramilitary Activities In and Against Nicaragua (Nicaragua v. United States of America)*, Merits, Judgment of 27 June 1986, General List No. 70.
- JACOBSEN Trudy, SAMPFORD Charles, THAKUR Ramesh (eds) (2016), *Re-envisioning Sovereignty: The End of Westphalia?*, Routledge.
- JAITNER Margarita Levin (2015), *Russian Information Warfare: Lessons from Ukraine*, in: Kenneth Geers (ed.), *Cyber War in Perspective: Russian Aggression Against Ukraine*, NATO CCDCOE Publications, Tallinn
- JOHNSON James (2019), *Artificial intelligence & future warfare: implications for international security*, "Defense & Security Analysis", vol. 35, no. 2, p. 147–169. DOI: 10.1080/14751798.2019.1600800
- KANIA Elsa B. (2019), *Chinese Military Innovation in the AI Revolution*, "The RUSI Journal", vol. 164, issue 5-6, p. 26–34. DOI: 10.1080/03071847.2019.1693803
- KAPLAN Jeffrey (2021), *A Conspiracy of Dunces: Good Americans vs. A Cabal of Satanic Pedophiles?*, "Terrorism and Political Violence", vol. 33, issue 5, p. 917–921. DOI: 10.1080/09546553.2021.1932342

- KLONOWSKA Klaudia (2020), Article 36: *Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare*, in: "Yearbook of International Humanitarian Law (YIHL)", vol. 23.
- LANG-IONATAMISHVILI Elina, SVETOKA Sanda (2015), *Strategic Communications and Social Media in the Russia Ukraine Conflict*, in: Kenneth Geers (ed.), *Cyber War in Perspective: Russian Aggression Against Ukraine*, NATO CCDCOE Publications, Tallinn
- LOOS Eugène, NIJENHUIS Jordy (2020), *Consuming Fake News: A Matter of Age? The Perception of Political Fake News Stories in Facebook Ads*, in: Q. Gao, J. Zhou (eds), *Human Aspects of IT for the Aged Population. Technology and Society. HCI 2020. Lecture Notes in Computer Science*, vol. 12209, Springer, Cham. DOI: 10.1007/978-3-030-50232-4_6
- MIZOKAMI Kyle (2020), *This Battlefield Tank Comes with an Xbox Controller*, "Popular Mechanics", <https://www.popularmechanics.com/military/weapons/a33457596/israeli-carmel-tank-video-games/> (29.07.2020)
- MORSTATTER Fred, SHAO Yunqiu, GALSTYAN Aram, KARUNASEKERA Shanika (2018), *From Alt-Right to Alt-Rechts: Twitter Analysis of the 2017 German Federal Election*, in: *WWW'18: Companion Proceedings of the The Web Conference 2018*, Lyon. DOI: 10.1145/3184558.3188733
- NADLER Anthony, CRAIN Matthew, DONOVAN Joan (2018), *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech*, Report, Data and Society Research Institute, <https://datasociety.net/library/weaponizing-the-digital-influence-machine/> (17.10.2018)
- NISSEN Thomas Elkjer (2015), *The Weaponization Of Social Media: Characteristics of Contemporary Conflicts*, Copenhagen.
- NOËL Jean-Christophe (2018), *Will artificial intelligence revolutionize the art of war?*, "Politique étrangère", issue 4, p. 159–170. DOI: 10.3917/pe.184.0159
- PARIS Roland (2020), *The Right to Dominate: How Old Ideas About Sovereignty Pose New Challenges for World Order*, "International Organization", vol. 74, issue 3, p. 453–489. DOI: 10.1017/S0020818320000077
- PAYNE Kenneth (2018), *Artificial Intelligence: A Revolution in Strategic Affairs?*, "Survival", vol. 60, issue 5, p.7–32. DOI: 10.1080/00396338.2018.1518374
- PFEFFER Jürgen, MAYER Katja, MORSTATTER Fred (2018), *Tampering with Twitter's Sample API*, "EPJ Data Science", vol. 7 (50). DOI: 10.1140/epjds/s13688-018-0178-0
- PIERRI Francesco, ARTONI Alessandro, CERI Stefano (2020), *Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections*, "PLoS ONE", vol. 15(1). DOI: 10.1371/journal.pone.0227821
- PRIER Jarred (2017), *Commanding the Trend: Social Media as Information Warfare*, "Strategic Studies Quarterly", vol. 11, no. 4, p. 50–85.
- REISMAN W. Michael (2003), *Assessing Claims to Revise the Laws of War*, "American Journal of International Law", vol. 97, issue 1, p. 82–90. DOI: <https://doi.org/10.2307/3087105>
- SCHMITT Michael N. (2013), *States and cyberspace*, in: *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Tallinn Manual draft), Cambridge. DOI: 10.1017/CBO9781139169288
- SCHMITT Michael N. (2014), *"Below the threshold" cyber operations: the countermeasures response option and international law*, "Virginia Journal of International Law", vol. 54 (3), p. 697–732.
- SCHMITT Michael N. (2017), *Law of international responsibility*, in: *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, Cambridge. DOI: 10.1017/9781316822524.010

- SHACKELFORD Scott J., RUSSELL Scott, KUEHN Andreas (2016), *Unpacking the International Law on Cybersecurity Due Diligence: Lessons from the Public and Private Sectors*, "Chicago Journal of International Law", vol. 17, no. 1.
- SILBERLING Amanda (2021), *Facebook will no longer allow advertisers to target political beliefs, religion, sexual orientation*, <https://techcrunch.com/2021/11/09/facebook-will-no-longer-allow-advertisers-to-target-political-beliefs-religion-sexual-orientation/> (10.11.2021)
- STATISTA (2019), *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions)*, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (30.04.2019)
- STATISTA (2021), *Number of monthly active Facebook users worldwide as of 2nd quarter 2021 (in millions)*, <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/> (31.10.2021).
- STINISSEN Jan (2015), *A Legal Framework for Cyber Operations in Ukraine*, in: Kenneth Geers (ed.), *Cyber War in Perspective: Russian Aggression Against Ukraine*, NATO CCDCOE Publications, Tallinn.
- U.S. DEPARTMENT OF THE TREASURY (2021), *Treasury Sanctions Russia with Sweeping New Sanctions Authority*, <https://home.treasury.gov/news/press-releases/jy0127> (15.04.2021)
- VARGO Chris J., GUO Lei, AMAZEEN Michelle A. (2018), *The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016*, "New Media & Society", vol. 20, issue 5, p. 2028–2049. DOI: 10.1177/1461444817712086.
- VOULGARI Iro et al. (2021), *Learn to Machine Learn: Designing a Game Based Approach for Teaching Machine Learning to Primary and Secondary Education Students*, in: *IDC '21: Interaction Design and Children*, Conference Proceedings, Athens. DOI: 10.1145/3459990.3465176
- WALSH Daniel Robert (2021), *Neutral Isn't Neutral: An Analysis of Misinformation and Sentiment in the Wake of the Capitol Riots*, MS Thesis, West Virginia University, DOI: 10.33915/etd.8055
- WAXMAN Matthew C. (2011), *Cyber-Attacks and the Use of Force: Back to the Future of Article 2(4)*, "Yale Journal of International Law", vol. 36. DOI: 10.2139/ssrn.1674565
- WHYTE Christopher (2020a), *Cyber conflict or democracy "hacked"? How cyber operations enhance information warfare*, "Journal of Cybersecurity", vol. 6, issue 1. DOI: 10.1093/cybsec/tyaa013
- WHYTE Christopher (2020b), *Deepfake news: AI-enabled disinformation as a multi-level public policy challenge*, "Journal of Cyber Policy", vol. 5, issue 2.