Chapter 4
Ethical Artificial Intelligence: An Approach to Evaluating Disembodied Autonomous Systems

Daniel Trusilo and Thomas Burri

In: Rain Liivoja and Ann Valjataga (eds), Autonomous Cyber Capabilities in International Law, forthcoming 2021, available on SSRN: https://ssrn.com/abstract=3816687

Abstract

Building off our prior work on the practical evaluation of autonomous robotic systems, this chapter discusses how an existing framework can be extended to apply to autonomous cyber systems. It is hoped that such a framework can inform pragmatic discussions of ethical and regulatory norms in a proactive way. Issues raised by autonomous systems in the physical and cyber realms are distinct; however, discussions about the norms and laws governing these two related manifestations of autonomy can and should inform one another. Therefore, this paper emphasizes the factors that distinguish autonomous systems in cyberspace, labeled *disembodied autonomous systems*, from systems that physically exist in the form of *embodied autonomous systems*. By highlighting the distinguishing factors of these two forms of autonomy, this paper informs the extension of our assessment tool to software systems, bringing us into the legal and ethical discussions of autonomy in cyberspace.

I. Introduction

'What our societies all over the world need is a shared and applicable ethical framework, to develop AI policies, regulations, technical standards, and business best practices.'[1] Addressing this call to action, our current project tackles the following question: How can an assessment tool designed to identify ethical issues of embodied autonomous systems be modified to apply to disembodied autonomous systems? The goal of such an undertaking is to inform a discussion about international norms and ethical principles that should apply to disembodied autonomous systems.

By applying an assessment tool we previously developed, henceforth referred to as the Schema, we are able to empirically identify ethical issues raised by autonomous disaster relief and weapon systems.[2]

---

[1] Luciano Floridi and Lord Tim Clement-Jones, 'The Five Principles Key to Any Ethical Framework for AI' (*New Statesman*, 20 March 2019) <https://tech.newstatesman.com/policy/ai-ethics-framework>.

[2] Markus Christen and others, 'An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications' (University of Zurich Digital Society Initiative White Paper no 1, 2017) <https://ssrn.com/abstract=3063617>.

Such systems necessarily have a physical manifestation. They are robots with a 'body' which is why we say that they are 'embodied'.[3] The scope of the Schema has so far been limited to such embodied systems. The first step in applying the Schema is to determine if an embodied system is autonomous. We make the determination according to a composite of: (1) autarchy, which in this context refers to a system's capacity to function independently from external energy sources, (2) independence of human control, (3) interaction with the environment, (4) learning, and (5) mobility.[4] This composite picture is how we define autonomy and determine if the Schema is in fact applicable to a given system. For the purpose of this chapter, robotic systems that meet the threshold of this composite definition of autonomy are referred to as *embodied autonomous systems* or simply embodied systems. We then evaluate a system according to thirty-seven aspects to determine potential areas of ethical concern. Our practical review of embodied autonomous systems using the Schema allows us to supplement the widely agreed upon framework of international humanitarian law, human rights law, and regulatory norms.

With the following discussion, we are advancing this research by extending the Schema to cover autonomous cyber operations, or software systems. Though software systems must be integrated with physical hardware to function, we are interested in exploring the idea of autonomy as it relates to algorithms that are created to function in their own right, not as code that controls a robotic system. We label such autonomous programs used in cyber operations as *disembodied autonomous systems*. A disembodied autonomous system, for the purposes of this chapter, is therefore a software program that demonstrates properties on a spectrum of a modified composite definition of autonomy, which will be discussed in greater detail in section three.

We have chosen to use the specific terminology of embodied and disembodied systems as these terms clarify our approach to the discussion about autonomy in cyberspace. They distinguish between the physical systems that are a combination of hardware and software, which we have experience evaluating, and software or algorithms that exist to carry out their own function. This distinction is mainly drawn for didactical purposes. The aim is to improve the Schema and extend its scope while furthering the discussion of autonomous systems and how ethically problematic aspects of such systems can be practically identified. Since embodied systems may incorporate elements of disembodied systems, and vice versa, it may not always be straightforward to distinguish the two. However, a neat and clean distinction may be unnecessary. If we manage to extend the scope of the Schema, making it

---

[3] For similar terminology, see Curtis EA Karnow, 'The Application of Traditional Tort Theory to Embodied Machine Intelligence' in Ryan Calo, Michael A Froomkin and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016).

[4] Christen and others (n 2).

comprehensive and inclusive of all autonomous systems, regardless of whether they are embodied or disembodied, then it will not matter whether the lines between the types of systems are blurred. There would simply be one Schema applicable to all autonomous systems.

The discussion is complicated by the fact that Artificial Intelligence ('AI') lies at the heart of the capabilities and capacities of the autonomous systems we are investigating but does not necessarily equate to autonomy in and of itself. While the relationship between autonomy and AI may have to be researched further,[5] it is our hope that the experience of researching and evaluating autonomy in embodied systems is transferrable to research concerning autonomy in cyberspace and therefore can add value to discussions surrounding what we have chosen to call disembodied autonomous systems.

We will first describe the urgent need to develop a method of identifying ethical issues related to the design and operation of disembodied autonomous systems.[6] Next, in order to determine how the Schema can be applied to systems that only exist in cyberspace, or disembodied autonomous systems, we highlight the factors that distinguish disembodied systems from embodied systems. We then explore and highlight those selected criteria of the Schema which will need to be modified in order to be applied to disembodied systems. The next section pushes the boundary further by discussing 'systems of systems', ie systems in which a collection of autonomous or semi-autonomous systems compose a larger system. This is particularly relevant in the discussion of cyber systems as the notion of a 'system' with a specific 'beginning' and 'end' becomes further blurred. We conclude with a brief overview of key takeaways from this chapter.

II. The Criticality of Evaluating Ethical Issues Raised by Disembodied Autonomous Systems

Developing a method to identify ethical issues concerning disembodied autonomous systems is practically relevant. Pure software programs with autonomous characteristics already exist. For example, in 2018 IBM Research demonstrated DeepLocker, an AI-powered malware that is able to evade detection until reaching a specific target. Using a deep neural network AI model, DeepLocker

5 See the discussion below, section II.

6 For further elaboration on the notion of autonomous cyber system see Tim McFarland, 'The Concept of Autonomy', this volume, ch 2.

seems benign, only deploying malicious code when it is triggered by its intended target, which it identifies through facial recognition, geolocation, and voice recognition.[7]

At the State level, US and Russian cyber operations have actively targeted each other's critical infrastructure, namely power grids.[8] On 13 March 2020, the cyberthreat to critical infrastructure was made palpable with an attack on Brno University Hospital in the Czech Republic, which led to the postponement of surgeries, the turning away of new patients, and the shutting down of all the hospital's computers.[9] The attack, coinciding with the global COVID-19 pandemic, demonstrates the life-threatening potential of cyberattacks and lends support to calls for an emergency regime for cyberspace.[10] It is not far-fetched to surmise that cyber weapons[11] being deployed by the US, Russia, and other actors may have autonomous capabilities, at least according to the composite definition of autonomy we apply. For example, the NotPetya cyber-attack of 2018 relied on self-propagating malware to become one of the most destructive and costly cyberattacks ever carried out.[12] Therefore, the concept of autonomy and what it means for disembodied systems must be discussed if any regime is to be relevant to current capabilities and trends.

Though there is an active debate about moral and legal issues related to autonomy in embodied weapon systems, or autonomous weapons systems, via the Convention on Certain Conventional Weapons, discussions concerning cyber systems have so far failed to address many of the similarly applicable

---

7 Marc Ph Stoecklin and others, 'DeepLocker: How AI Can Power a Stealthy New Breed of Malware' (*Security Intelligence*, 8 August 2018) <https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/>.

8 A June 2019 article in the New York Times publicized the years-long cyber operations by both Russian and US entities to implant malicious code in their adversary's critical infrastructure. David E Sanger and Nicole Perlroth, 'US Escalates Online Attacks on Russia's Power Grid' (*New York Times*, 17 June 2019) <https://www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html>.

9 Matt Burgess, 'Hackers are Targeting Hospitals Crippled by Coronavirus' (*Wired*, 22 March 2020) <https://www.wired.co.uk/article/coronavirus-hackers-cybercrime-phishing>.

10 See Henning Lahmann's blog post calling for an emergency regime related to cyberattacks on hospital infrastructure. Henning Lahmann, 'Cyberattacks against Hospitals during a Pandemic and the Case for an Emergency Regime for Cyberspace' (*Fifteen Eightyfour*, 20 April 2020) <http://www.cambridgeblog.org/2020/04/cyberattacks-against-hospitals-during-a-pandemic-and-the-case-for-an-emergency-regime-for-cyberspace/>.

11 A broad definition of a cyber weapon includes software and IT systems that, through ICT networks, manipulate, deny, disrupt, degrade or destroy targeted information systems or networks. The pros and cons of this definition is discussed in Tom Uren, Bart Hogeveen and Fergus Hanson, 'Defining Offensive Cyber Capabilities' (Australian Strategic Policy Institute, 4 July 2018) <https://www.aspi.org.au/report/defining-offensive-cyber-capabilities>.

12 Andy Greenberg, 'The Untold Story of NotPetya, the Most Devastating Cyberattack in History' (*Wired*, 22 August 2018) <https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/>.

implications of autonomy.[13] This situation may be partly due to the tendency to silo discussions of legal ramifications of various new technologies in warfare through a technology-specific approach.[14] However, this tendency is alarming considering the likelihood that decision authorities will be delegated to both embodied and disembodied autonomous systems and that the various systems are conflated when the discussion centers on autonomy.

In a 2016 interview with the Washington Post, the then US Deputy Secretary of Defense, Robert Work, stated that the use of unmanned systems by the US Department of Defense (DoD) is inexorable. In clarifying DoD's position, Work explained that autonomy is a matter of delegating authorities to unmanned systems in a battle network and that delegation of authority can be expected in situations in which machines have faster than human reaction times. Work then specifically identified electronic and cyber warfare as examples of situations in which machines have faster than human reaction times, warranting the delegation of decision making authorities to unmanned systems.[15] Despite this recognition, the US DoD Directive 3000.09 on Autonomy in Weapons Systems, which addresses authorities related to autonomous systems, explicitly states that it does not apply to autonomous or semi-autonomous systems for cyberspace operations.[16] A United Nations Institute for Disarmament Research (UNIDIR) paper on Autonomous Weapon Systems (AWSs) and Cyber Operations notes that the DoD directive excluded cyber considerations for pragmatic reasons – the directive was urgently needed and addressing autonomy in cyber operations would have delayed publication of the directive.[17]

---

[13] For a discussion of the state of international law and autonomous cyber operations as well as the importance of addressing autonomous cyber capabilities, see Rain Liivoja, Maarja Naagel and Ann Väljataga, 'Autonomous Cyber Capabilities under International Law' (NATO CCDCOE 2019) <https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/>.

[14] Rain Liivoja, 'Technological Change and the Evolution of the Law of War' (2015) 97(900) International Review of the Red Cross 1157.

[15] See interview with US Deputy Secretary of Defense Robert Work in 'David Ignatius and Pentagon's Robert Work Talk about New Technologies to Deter War' (*The Washington Post*, 31 March 2016) <https://www.washingtonpost.com/video/postlive/david-ignatius-and-pentagons-robert-work-on-efforts-to-defeat-isis-latest-tools-in-defense/2016/03/30/0fd7679e-f68f-11e5-958d-d038dac6e718_video.html>. The referenced discussion concerning autonomy and the delegation of authorities begins at 27:18.

[16] US Department of Defense, Directive 3000.09: Autonomy in Weapons Systems (21 November 2012, incorporating change 1, 8 May 2017) < https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>.

[17] United Nations Institute for Disarmament Research, 'The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations' (16 November 2017) <https://www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber>.

The DoD Directive 3000.09 was published in 2012 and updated in 2017, yet autonomy in cyber operations remains unaddressed.

The UNIDIR Report highlights the fact that international discussions related to what we call embodied and disembodied systems are completely divorced from each other 'with virtually no overlap between the participating experts and policy practitioners', despite the relevance of autonomy for both.[18] The Group of Governmental Experts ('GGE') discussions related to cyber security have addressed neither the concept of meaningful human control nor Article 36 obligations on the testing of the means and methods of cyber warfare,[19] both of which are topics that are heavily featured in GGE discussions of embodied AWSs.

In a similar vein, there is a need to bridge the discussions of AI and autonomy. The 2019 Defense Innovation Board's (DIB) *Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* were adopted as principles by the DoD on 24 February 2020.[20] Bounding the applicability of the DIB's recommendations to AI, the report explicitly states 'AI is not the same thing as autonomy.'[21] The report goes on to highlight that DoD Directive 3000.09, 'neither addresses AI as such nor AI capabilities not pertaining to weapon systems.'[22] Though it is clear that AI is not the same thing as malware, the fact remains that AI may be used as part of malware and cyber operations in general.[23] Taken as a whole, this information signals the gap in the framing of ethical and legal discussions surrounding the subjects of AI and autonomy despite their convergence in disembodied autonomous systems.[24]

---

[18] ibid.

[19] James Lewis and Kerstin Vignard, 'Report of the International Security Cyber Issues Workshop Series' (United Nations Institute for Disarmament Research, 2016) <https://www.unidir.org/files/publications/pdfs/report-of-the-international-security-cyber-issues-workshop-series-en-656.pdf>.

[20] US Department of Defense, 'DOD Adopts Ethical Principles for Artificial Intelligence' (24 February 2020) <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[21] Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artifiical Intelligence by the Department of Defense' (US Department of Defense, 31 October 2019) <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF>.

[22] ibid.

[23] Stoecklin (n 7).

[24] Compare Heather M Roff, 'Artificial Intelligence: Power to the People' (2019) 33 Ethics & International Affairs (2) 127, 140: '[W]e need to be careful of conflating AI with automation or autonomy, for doing so risks aggregating benefits and harms in different ways, when we would do better to keep them separate.' On autonomy and AI, see also Alan L Schuller, 'At the Crossroads of Control: The Intersection of Artificial

Article 36 of the 1977 Additional Protocol to the 1949 Geneva Conventions requires States to conduct a weapon review prior to the acquisition or adoption of 'a new weapon, means or method of warfare.'[25] There is an existing body of literature on how the design and testing process applies to non-autonomous cyber weapons.[26] But there are further challenges to applying an Article 36 review to a system that has autonomous capabilities.[27] These challenges have led to an active debate about how to apply weapons reviews to embodied AWSs. However, autonomy does not figure in the review of cyber weapons, meaning cyber weapons that are currently under development are being designed and tested without any specific institutionalized rules or international norms. This is problematic as autonomous cyber weapons that incorporate learning, even if such learning is frozen at the moment of operationalization, may behave unpredictably.[28] The complications are obvious when one looks at the commentary to Rule 110 of the *Tallinn Manual 2.0*, in which the consensus of the group of governmental experts states: 'Any significant changes to means or methods necessitate a new legal review.'[29] Based on this language, a State that deploys an autonomous cyber weapon may no longer be in compliance with the law of armed conflict once the autonomous cyber weapon goes beyond predicted behavior or learns and modifies it. Therefore, addressing autonomy when ethically evaluating such systems is vitally important.

Intelligence and Autonomous Weapons Systems with International Humanitarian Law' (2017) 8 Harvard National Security Journal (2) 379, 390 ff.

[25] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 ('AP I') art 36.

[26] The *Tallinn Manual 2.0*, a study on the application of international law to cyber-warfare, includes part IV on Cyber Armed Conflict with extensive rules concerning the means and methods of warfare and specific guidance on the applicability of the Article 36 weapons review process to cyber weapons. Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

[27] Vincent Boulanin and Maaike Verbruggen, 'Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies' (Stockholm International Peace Research Institute, 2017) <https://www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf>.

[28] United Nations Institute for Disarmament Research (n 17).

[29] *Tallinn Manual 2.0* (n 26) commentary to Rule 110, [9].

III. Distinguishing Disembodied Autonomous Systems from Embodied Autonomous Systems

In order to identify how the regulatory framework applicable to embodied systems can be applied to disembodied systems and to extend our tool, the emphasis must be placed on what *distinguishes* disembodied autonomous systems from embodied autonomous systems.

When focusing on disembodied autonomous systems the task of determining the constitutive parts of the 'system' to be assessed changes. With embodied systems, the existence of some kind of robotic manifestation imparts an intuition of where and how the system is bounded. This intuition is less clear with regard to disembodied systems because of their lack of a physical manifestation. Disembodied systems may also propagate without incurring additional cost; they can be passed on like fire.[30] Such an analogy allows one to envision a disembodied autonomous system spreading widely. Such a possibility may warrant even more vigilance in the development of disembodied autonomous systems based on the precautionary principle.[31]

The autonomy of an embodied system may be viewed as *composite*. In this way, a system's autonomy may be assessed according to the five axes of: (1) autarchy, (2) independence of human control, (3) interaction with the environment, (4) learning, and (5) mobility. A system can then be positioned on each axis resulting in an overall picture of its autonomy.[32] For disembodied systems, however, 'autarchy' becomes meaningless. Electricity is a pre-condition for software to function so if the environment that a disembodied system inhabits is functioning, no additional battery or fuel source is required for the disembodied system to also function. The concept of 'mobility' also changes when applied to a disembodied system as software is incapable of physically moving on its own, though it may migrate through a network. Therefore, the concept of mobility is not applicable if meant in the physical sense and must be adapted to relate to a disembodied system's characteristics. Applying the above considerations, an autonomous disembodied system must be one that, to a certain degree, can:

---

[30] We draw here on a statement made by a legal scholar with regard to legal personhood: 'legal personhood is like fire: it can be granted by anyone who already has it'. Shawn Bayern, 'The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems' (2016) 2 European Journal of Risk Regulation 297, 304.

[31] See AP I art 57; Jonathan David Herbach, 'Into the Caves of Steel: Precaution, Cognition and Robotic Weapons Systems Under the International Law of Armed Conflict' (2012) 4(3) Amsterdam Law Forum 3, 6 ff.

[32] Christen and others (n 2) 10.

(1) operate independent of human control once deployed, (2) interact with its environment based on characteristics that define the environment, and (3) learn.

A system may vary along the described axes, that is to say, not every system needs to be capable of learning. To a certain extent, we accept Tim McFarland's statement that independence from operator control is not an ideal determinant of autonomy.[33] However, we construe the term in a similar way that McFarland interprets autonomy. This means that typically a programmer/operator defines the high-level goal to be achieved by the autonomous system, while the low-level steps are subject to the system's 'discretion'[34] – it being understood that low-level steps also need to be programmed or at least learned at one point. This construction of 'independence of operator control' also has the advantage of focusing the Schema on systems exhibiting a certain degree of complexity, while excluding simple software. Image processing software such as Adobe Photoshop, for instance, is not programmed to attain high-level goals and hence cannot be considered to be 'independent from control', even though it can remove red eyes at the click of a button.

The requirement that a system interacts with the environment mirrors McFarland's emphasis on the environmental uncertainty that autonomous systems typically have to cope with. However, the Schema does not insist that the environment be uncertain. This would initially have been conceivable when the focus of the Schema had been on embodied systems. For embodied systems, uncertainty of physical environmental factors are typically a hard to overcome challenge. The pathway ahead of a robot may, for instance, become icy, there may be debris, or gusts of wind may unexpectedly impact it. In cyberspace in contrast, to require a system to cope with environmental uncertainty seems to go beyond what is necessary. The environment in cyberspace is more structured and therefore the types of interactions a system can possibly face are more limited. Ice, rubble, and wind cannot occur in cyberspace (except in metaphors). So, if the Schema should also cover disembodied systems, requiring

---

[33] McFarland (n 6) [35]: 'Regardless of its complexity, autonomous software amounts to a set of instructions guiding a system toward such a goal. Those instructions may endow a system with a capacity for complex actions and responses, including the ability to operate effectively in response to new information encountered during operations which may not be precisely foreseeable to a human operator. However, that does not constitute independence from human control in any sense. Rather, it is best seen as control applied in a different way, in advance rather than in real time.' And at [52]: '"autonomy", as the concept applies to software, does not mean freedom from of human control; it is, rather, a form of control.'

[34] ibid [9]. Unlike McFarland, we refrain from using terms like 'intent' and 'awareness' to avoid the risk of anthropmorphizing the system. Cf Neil M Richards and William D Smart, 'How Should the law think about robots?' in Ryan Calo, Michael A Froomkin, and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016) 13: '[W]hen it comes to new technologies, applying the right metaphor for the new technology is especially important. How we regulate robots will depend on the metaphors we use to think about them. There are multiple competing metaphors for different kinds of robots, and getting the metaphors right will have tremendously important consequences for the success or failure of the inevitable law (or laws) of robotics.'

a capacity to cope with environmental uncertainty would be unnecessary. Indeed, if coherently applied across embodied and disembodied systems, such a requirement may prove overly exclusionary.

There is a clear distinction between embodied systems that are *intended to cause harm* ('weapons') and systems that are not intended to cause harm (for example, search and rescue systems). In the case of the former, the Schema evaluates an additional set of criteria. While the distinction may also make sense for disembodied systems, the notion of 'harm' may have to be construed more broadly. Observational systems, such as systems that exclusively serve to gather data, may be considered not to cause harm. On the other hand, not only systems that cause physical damage (by kinetic means, for instance breaking infrastructure), but also systems that actively cause malfunctions, delay services, and so on, may be considered harmful when such malfunctions or delays can lead to actual physical harm.

The *Tallinn Manual* provides useful orientation on the notion of harm. Regarding the definition of the use of force, the majority of the international group of experts agreed, 'acts that injure or kill persons or damage or destroy objects are unambiguously uses of force'.[35] The 2010 Stuxnet cyber-attack on the Iranian nuclear program that led to the destruction of centrifuges is an oft-cited example of a real-world cyber operation that resulted in physical damage. Such an attack could be considered a use of force. Furthermore, the consensus view of the experts, in commentary to Rule 13 of the Tallinn Manual, was that the aggregate sum of a series of cyber-attacks can be treated as a composite armed attack thus allowing a State to exercise the right of self-defense.[36] A disembodied autonomous system, not explicitly designed to cause physical damage, may spread through a network where it was not intended to operate, causing physical damage or delays in service to multiple systems that then lead to physical damage and/or the loss of life. Therefore, to apply our assessment tool to disembodied systems, we must revisit our method of determining if a system is intended to cause harm.[37] For disembodied systems, both intention and harm should notably be understood in a less direct sense. When the operation of a disembodied system may indirectly lead to harm, we will have to apply the set of criteria that are normally reserved for weapon systems in order to ensure the ethical implications of operating the evaluated system are fully considered.

---

35 *Tallinn Manual 2.0* (n 26) commentary to rule 11 [8].

36 ibid commentary to rule 13 [8].

37 The notion of 'harm', which we have chosen to employ in order to be inclusive of all potentially ethically problematic systems, is distinctly different than the notion of an 'attack'. As pointed out by Rain Liivoja and Tim McCormack, the question of what kind of cyber operations could trigger armed conflict while falling below the threshold of an attack is not thoroughly addressed in the Tallinn Manual. Rain Liivoja and Tim McCormack, 'Law in the Virtual Battlespace: The Tallin Manual and the Jus in Bello' (2012) 15 Yearbook of International Humanitarian Law 45.

IV. The Evaluation of Disembodied Autonomous Systems

Once a disembodied system has been determined to have aspects of autonomy and its potential to cause harm is known, we can apply the Schema's criteria to identify ethical issues raised by a particular system. The majority of criteria that are applicable to evaluating an embodied system will directly cross-over to an evaluation of a disembodied system. An example of a directly relatable criteria is the concept of 'emergent properties.' The question of whether a system to system interaction can yield unexpected or emergent properties is relevant, but it needs no modification to apply to a disembodied system. For the purposes of this chapter we will not highlight criteria that are directly transferrable but rather the criteria that must be modified or interpreted differently to account for differences between embodied and disembodied autonomous systems.

One criterion, classified in our tool as an aspect of how the system interacts with the operator, which requires review and re-interpretation is 'responsibility attribution'. Whereas an embodied system is a physical entity that an operator must deploy from a specific location, disembodied systems are less tied to physicality and location. They migrate through the network of fiber-optic cables that connect the globe and 'lend themselves to plausible deniability'.[38] Furthermore, embodied systems are physically constituted of manufactured components, which can be serial-numbered and traced. A disembodied system is a sequence of code and may be hidden within a completely innocuous program that comes from another source. For these reasons, tracing an autonomous cyber weapon for attribution purposes, even with an array of digital forensic tools, may prove time and resource intensive, if not nearly impossible.

Further complicating the ability to attribute a system to a responsible party and raising questions about proliferation, is the possibility of a disembodied system multiplying ('self-replicating') without any immediate command to do so by the human that initially developed and programmed the system. This possibility raises questions of not only how international actors can identify the human party that is responsible for the actions of a disembodied system but also if the distribution and proliferation of such a system could be monitored even if international regulations were agreed upon. Lastly, in a chapter exploring the human element of cyber operations, David Danks and Joseph H Danks emphasize the challenge of clear responsibility attribution even if it is technically known who initially programmed a

[38] In a July 2019 *New Yorker* article, Sue Halpern describes the June 2019 use of cyber weapons by the US against Iran in retaliation for the downing of a US surveillance drone. The article frames the challenge of attribution as a question, asking, 'How do you levy a threat when it's not clear where an attack is coming from or who is responsible?' See Sue Halpern, 'How Cyber Weapons are Changing the Landscape of Modern Warfare' (*The New Yorker*, 18 July 2019) <https://www.newyorker.com/tech/annals-of-technology/how-cyber-weapons-are-changing-the-landscape-of-modern-warfare>.

system as the speed and velocity of cyber-actions means that humans will inevitably be out-of-the-loop when events occur.[39] These questions echo the notion of a responsibility gap, a well-known concern with embodied autonomous systems.[40]

Considering the deployment conditions of a system, we also need to modify the criterion that assesses a system's 'effects on [the] general population'. When evaluating an embodied system one can determine if the system is likely to come into contact with a civilian population such as crowds and other neutral populations. Disembodied systems, on the other hand, may come into contact and influence a population without the individuals ever knowing they were interacting with the systems. For instance, though humans controlled the operations, Cambridge Analytica was able to use AI to aggregate vast amounts of data and deploy targeted disinformation campaigns to influence unwitting voters via social media and affect democratic elections.[41] We will therefore explore, in the following section, how to assess the potential impact of a disembodied autonomous system that is designed to be deployed in an interconnected network of both military and civilian infrastructure.

A physical characteristic of an embodied system that can be assessed fairly easily is the 'degree of lethality'. One can determine the properties of an embodied system's physical armaments – are they lethal or not? By definition, however, a disembodied system will have no physical weapons, though it may be designed in such a way as to make it lethal. For example, a cyber weapon that targets a self-driving automobile's operating system and then causes the vehicle to accelerate into pedestrians, is lethal. What core questions must be asked then to determine the disembodied system's intended use and its degree of lethality beyond the notion of physical armaments?

Another criterion, classified in the Schema as a behavioral characteristic, relates to 'constraining the system in time and space'. An embodied system may be temporally and geographically bound through a variety of methods. However, a disembodied system does not operate in a physical space. That is not to say that constraints cannot be applied to disembodied systems or that such constraints cannot be

---

[39] David Danks and Joseph H Danks, 'Beyond Machines: Humans in Cyber Operations, Espionage, and Conflict' in Fritz Allhoff, Adam Henschke, and Bradley Jay Strawser (eds), *Binary Bullets: The Ethics of Cyberwarfare* (Oxford University Press 2016).

[40] Robert Sparrow, 'Killer Robots' (2007) 24(1) Journal of Applied Philosophy 62.

[41] See US Senate Select Committee on Intelligence, 'Report of the US Senate Select Committee on Intelligence: Russian Active Measures Campaigns and Interferences in the 2016 US Election, Volume 2: Russia's Use of Social Media with Additional Views' (116th Congress, Report 116-XX, 2019) <https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf>; European Commission, 'Code of Practice on Disinformation' (26 September 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

assessed, but the concept of boundaries will need to be modified to account for the non-physical environment of cyber space.

When evaluating the behavioral characteristics of a system, one must also be able to guarantee the reliability of the system's behavior. Relating this to the assessment of a system's 'targeting' capability, one must be able to say with confidence that a system will reliably target what it has been deployed to target. When applied to an embodied system, one can determine if the system is able to reliably distinguish between lawful and unlawful targets via extensive testing including an Article 36 weapons review. Even if Article 36 reviews of physical weapons are carried out (they are not always), testing complex autonomous embodied systems that have limited autarchy, mobility, and cannot self-replicate, is already difficult.

Applying the requirement of reliability to a disembodied system presents a challenge of a different order of magnitude.[42] The Tallinn Manual explicitly requires certainty that both offensive and defensive cyber-attacks are directed at lawful targets.[43] However, the Tallinn Manual does not take into account autonomy. Autonomous cyber weapons may be deployed in a lawful manner, but if a disembodied system has the ability to choose targets by means of AI, how can one ensure the system's targets will remain lawful? This question is especially difficult to address given the '(in)ability to predict rapid sequences of events that can result from the use of automated responses (the chain reaction challenge).'[44] This chain reaction challenge means that the behavior of autonomous cyber systems could result in an unpredictable escalation of consequences through feedback loops that are too fast for a human to stop.[45]

Other assessment criteria in the Schema may be irrelevant to non-physical entities. In the modified Schema that applies to disembodied autonomous systems these criteria can simply be ignored. They

---

[42] Robert Work stated in a 2016 interview with author Paul Scharre: 'When you delegate authority to a machine, it's got to be repeatable… So, what is going to be our test and evaluation regime for these smarter and smarter weapons to make sure that the weapon stays within the parameters of what we expect it to do?' Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (WW Norton 2018) 149.

[43] Jeffrey S Caso, 'The Rules of Engagement for Cyber-Warfare and the Tallinn Manual: A Case Study' in *The 4th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER)* (IEEE 2014).

[44] David Danks and Joseph H Danks, 'The Moral Permissibility of Automated Responses During Cyberwarfare' (3013) 12(1) Journal of Military Ethics 18, 19.

[45] For a discussion of human control over (embodied) autonomous weapon systems by means of veto power, see Noel Sharkey, 'Staying in the Loop: Human Supervisory Control of Weapons' in Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß (eds) *Autonomous Weapons Systems – Law, Ethics, Policy* (Cambridge University Press 2016) 35–6.

include: the appearance of the system, physical safeguards, and environmental effects. A detailed discussion of each of these criteria is not warranted here.

V. The Near Future Challenge: Evaluating a System of Systems

Near-future applications of autonomous systems may incorporate networks of interconnected disembodied and embodied systems. From an operational perspective, intelligent collective behavior, or swarm strategies, offer incredible promise of new and powerful capabilities.[46] This concept of networked autonomous systems, or *systems of systems*, complicates attempts at classification and evaluation.[47] Evaluating a system of systems according to norms, values, and regulations may not be the same as individually evaluating its constituent parts. While we are discussing the broadening of the Schema to disembodied systems, let us therefore briefly contemplate the implications of an aggregate of systems in which both embodied and disembodied systems may play a role.

As Paul Scharre highlighted in his 2018 book, *Army of None*, networked systems will have the capability to perform some tasks independently with human oversight, 'particularly when speed is an advantage… Future weapons will be more intelligent and cooperative, swarming adversaries.'[48] With increasingly advanced defensive weapons, the use of swarms of low-cost unmanned systems is likely. Such systems are considered attritable, meaning a force can plan on losing any number of the individual systems without detrimental consequences to strategic outcomes, budgets, or overall capabilities. Such swarms are being developed for ground and sea operations and have already been operationally deployed in air operations.

On 1 March 2020, the Turkish military announced it had deployed swarms of drones to attack Syrian government forces.[49] Though the extent of the autonomy of the systems deployed is not evident, the

---

[46] Joe Burton and Simona R Soare, 'Understanding the Strategic Implications of the Weaponization of Artifical Intelligence' in Tomáš Minárik and others (eds), *2019 11th International Conference on Cyber Conflict (CyCon)* (NATO CCDCOE 2019).

[47] According to Airbus, '[t]he cornerstone of FCAS is the next-generation weapon system where next-generation fighters team up with remote carriers as force multipliers. Additionally, manned and unmanned platforms also will provide their uniqueness to the collective capabilities while being fully interoperable with allied forces across domains from land to cyber. The air combat cloud will enable the leveraging of networked capabilities of all pooled platforms.' Airbus, 'Future Combat Air System (FCAS)' (2020) <https://www.airbus.com/defence/fcas.html>.

[48] Scharre (n 42) 93.

[49] Selcan Hacaoglu, 'Turkey's Killer Drone Swarm Poses Syria Air Challenge to Putin' (*Bloomberg*, 1 March 2020) <https://www.bloomberg.com/news/articles/2020-03-01/turkey-s-killer-drone-swarm-poses-syria-air-challenge-to-putin>.

Turkish military does have weaponized drones that are capable of automated functions.[50] The March 2020 operation is the first instance of a government explicitly stating it had used a swarm of weaponized drones in a coordinated offensive.

Turkey is not the only State racing to develop low-cost, AI piloted and networked weapon platforms. Manned fifth-generation aircraft like the F-35 Joint Strike Fighter have production costs close to USD 100 million per aircraft. To augment expensive, low-volume platforms, lower-cost, autonomous aircraft such as the XQ-58 Valkyrie are being prototyped. For example, Assistant Secretary of the US Air Force Will Roper stated that the US is developing a program known as Skyborg in order to prototype an AI piloted wingman capability. The publicly stated goal of the Skyborg program is to have autonomous and attritable systems ready by 2023.[51]

These facts reinforce the urgent need for a method of ethically evaluating not just individual, embodied and disembodied autonomous systems but rather the combined effect of a network of autonomous systems coordinated and controlled by AI as a system of systems. Though one could classify a swarm of drones as one whole embodied autonomous system and apply the Schema as it is, the lines begin to blur when autonomous cyber systems play a role in coordination with embodied autonomous systems. For example, it is conceivable that an autonomous software platform that only exists in cyberspace could be used to command and control an interconnected fleet of systems including drone swarms; associated logistical support systems; and intelligence, surveillance, and reconnaissance assets.[52] The notion of a system of systems based entirely in cyberspace also raises questions. For example, how could one determine where in a cyber system of systems the 'system' to be evaluated begins or ends?

VI. Conclusion

Certain norms purportedly govern cyberspace, but the kind of consensus supporting traditional law has so far proven elusive. Tools, such as a modified version of the Schema, which can be used to identify ethical issues raised by disembodied autonomous systems, must be further developed. Such tools can

---

[50] See Baykar Defence technical description of the Bayraktar TB2 unmanned aerial vehicle and its capabilities including fully autonomous taxiing, take-off, landing, and cruise. Baykar, 'Bayraktar TB2' (2019) <https://baykardefence.com/uav-15.html>.

[51] Valerie Insinna, 'Under Skyborg Program, F-35 and F-15EX Jets could Control Drone Sidekicks' (*Defense News*, 22 May 2019) <https://www.defensenews.com/air/2019/05/22/under-skyborg-program-f-35-and-f-15ex-jets-could-control-drone-sidekicks/>.

[52] For an example of a currently operational system of systems that uses a software platform to command and control multiple interlinked hardware systems see the US Navy's Aegis Combat System. The Aegis system features fully-autonomous and semi-autonomous functions. Lockheed Martin, 'Aegis: The Shield of the Fleet' (2 May 2020) <https://www.lockheedmartin.com/en-us/products/aegis-combat-system.html>.

inform pragmatic discussions of ethical and regulatory norms in a proactive way. There is a clear link between the issues raised by autonomous embodied systems and those raised by disembodied autonomous systems. Discussions about the norms and laws governing these two distinct yet related manifestations of autonomy should inform one another. Already much of our research focuses on how the underlying software used in autonomous robotics manifests in the physical world. The extension of our assessment tool to software systems is an attempt at linking the two discussions and brings us into the legal and ethical discussions of cyberspace.

About the authors: Thomas Burri is a professor of international law and European law at the University of St. Gallen in Switzerland. His research investigating AI and autonomous systems for almost a decade has been widely published. Daniel Trusilo is a PhD student at the University of St. Gallen. He previously served as a member of the U.S. armed forces, *inter alia* in Iraq. Most recently he worked as a humanitarian advisor to the military for the U.S. Agency for International Development.