2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Autonomous Decision-Making Processes and the Responsible Cyber Commander

Jody M. Prescott
Senior Fellow, West Point Center for the Rule of Law
Adjunct Professor, Department of Political Science, University of Vermont
Burlington, Vermont, USA
jody.prescott@us.army.mil; 01jpresc@uvm.edu

Abstract: With cyber operations conceivably moving at near light speed, commanders in cyber warfare will likely need to rely extensively upon autonomous decision-making processes (ADPs) to be effective. For commanders to meet their obligations under the Law of Armed Conflict (LOAC) and complementary Rules of Engagement (ROE), these ADPs must function in a manner compliant with both. To better understand how such ADPs might be effectively used, it is important to consider the operational challenges cyber commanders face in conducting cyber warfare, the different options available to cyber commanders to decrease the time frame required for making effective, LOAC-compliant decisions, and how ethical ADPs might be created. To that end, this paper focuses on the development of programme architecture and procedures that will be necessary to meet LOAC and ROE requirements, rather than the applicable law or potential ROE.

Keywords: *law of armed conflict, commander, autonomous decision-making processes*

1. INTRODUCTION

From the perspective of the law of armed conflict (LOAC), a cyber commander's battle-space resembles a game of three-dimensional chess in important respects. On the level in which the effects of cyber operations might ripple into the geophysical world and injuries to people and damage to tangible objects may be reasonably foreseen, LOAC will likely apply just as it does during traditional kinetic warfare.[1] This would include criminal responsibility for a cyber commander whose actions or inaction resulted in LOAC violations.[2] On a second level, the one in which the effects of cyber actions remain in cyberspace and result at most in the manipulation or deletion of non-critical data, more cyber annoyance than cyber attack, LOAC would not appear to apply at all.[3] Instead, the governing authorities are likely national rules of engagement (ROE). There is likely a third level in between these two; the one in which the direct effects of cyber actions remain in cyberspace but very serious indirect effects register in the geophysical world as a result of the degradation or destruction of critical national cyber infrastructure. U.S. cyber strategy documents[4] and statements of Department of Defense (DOD) officials[5] suggest that LOAC-like principles embedded in ROE might be part of the decision calculus governing whether and how to respond to such cyber activities.

Thus, although each level is in play simultaneously, unlike a game of three-dimensional chess, different rules apply to each level. Further, unlike the deliberative pace of chess, the operational tempo of cyberspace action is much more intense and capable of moving at almost the speed of light.[6] As a matter of operational necessity, cyber commanders will need to rely extensively upon autonomous decision-making processes (ADPs) that conduct cyber response activities and operations as the

---

1 Tallinn Manual on the Law of Cyber Warfare, Rule 13, para. 6, 55; Rule 29, para. 1, at 91; Rule 30, para. 5, 93.

2 *Id*., Rule 24, at 80.

3 *See id*., Rule 13, para. 6, 55 ("acts of cyber intelligence gathering and cyber theft, as well as cyber operations that involve brief or periodic interruption of non-essential cyber services, do not qualify as an armed attack.").

4 Department of Defense, Cyber Policy Report Pursuant to Section 934 of the NDAA of FY 2011 (Nov. 2011), 3-8, *available at* http://www.defense.gov/home/features/2011/0411_cyberstrategy/docs/NDAA%20Section%20934%20Report_For%20webpage.pdf [hereinafter "Cyber Report"].

5 *See* Ellen Nakashima, *Pentagon proposes more robust role for its cyber-specialists*, washingtonpost.com (Aug. 9, 2012), *available at* http://www.washingtonpost.com/world/national-security/pentagon-proposes-more-robust-role-for-its-cyber-specialists/2012/08/09/1e3478ca-db15-11e1-9745-d9ae6098d493_story.html; Amber Corrin, *Cyber warfare: New Battlefield, new rules*, FCW.com, Jul. 9, 2012, *available at* http://fcw.com/articles/2012/07/15/feat-inside-dod-cyber-warfare-rules-of-engagement.aspx; William J. Lynn, Remarks on the Department of Defense Cyber Strategy, speech made in Washington, D.C., (July 14, 2011) *available at* http://www.defense.gov/Speeches/Speech.aspx?SpeechID=1593.

6 Thomas C. Wingfield et al, *Optimizing Lawful Response to Cyber Intrusions*, 2 (2005) (unpublished paper), *available at* http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA464203 .

product of computer-driven decision cycles lasting perhaps no more than fractions of a second. There appear to be misgivings in general about the use of ADPs when human targets are involved.[7] From a strictly operational perspective, however, the risk of not utilising ADPs to conduct cyber operations might be unacceptable, and the temptation to accelerate the pace of decision-making by reducing the role of the human commander might be very strong.[8] As defensive and offensive cyber measures become more sophisticated, the demarcation between the two might become more blurred,[9] and the issues regarding the propriety of these different uses of force more intertwined.

To better understand how cyber commanders might use ADPs in a LOAC-compliant manner, this article first identifies the operational challenges faced by current cyber commanders. Second, the different options available to compress the time frames within which cyber commanders must make their decisions are explored, and the relative advantages and disadvantages of each assessed. Third, on-going research in a related field – the development of autonomous geophysical robot weapons – is then examined to highlight the challenges involved in seeking to embed LOAC and ROE principles, prohibitions and permissions into cyber ADPs. In conclusion, this paper suggests that in combination with other options to compress a cyber commander's decision timeframe, LOAC- and ROE-compliant ADPs could constitute an effective and necessary means by which the obligation of command responsibility is met in cyber operations, and that steps should be taken immediately to ensure that LOAC principles are incorporated into ADP design processes.

2. OPERATIONAL CHALLENGES FACING CYBER COMMANDERS

In the most comprehensive study of its kind available in the public domain to date, a senior U.S. naval officer surveyed a number of senior U.S. officers who had cyber operations experience and who served on the staff of the U.S. Chairman of the Joint

---

[7] *See* Ronald C. Arkin, Governing Lethal Behavior in Autonomous Robots, 52-55 (2009) (in a 2008 survey of 430 roboticists, military personnel, members of the general public, and policy makers, over half of the respondents found the taking of human life by an autonomous robot unacceptable). 66% of those surveyed believed that the robot should be held to higher ethical standards than soldiers. *Id.* at 55.

[8] Thomas K. Adams, *Future Warfare and the Decline of Human Decisionmaking*, 31 Parameters 57, 66 (Winter 2001/2002).

[9] Christopher Ford, *Cyber Operations: Some Policy Challenges*, newparadigmsforum.org (June 3, 2010), *available at* http://www.newparadigmsforum.com/NPFtestsite/?p=270 (last visited Dec. 16, 2012).

Chiefs of Staff. The survey included 15 men and six women,[10] who had on average 22 years of military service, over one year of cyber warfare decision-making experience, and almost two and one-half years' of service on the Joint Staff.[11] All of these officers had master's degrees, over half held two or more master's degrees, and 14% had professional degrees.[12] On the basis of their experiences, the study participants identified several significant concerns they had with conducting cyber operations. Their concerns included the uncertainty that results from the complexity of cyber operation response processes, the technical challenges in discriminating between military objectives and civilian objects, the difficulty in applying LOAC and ROE to cyber operations, and importantly for purposes of this article, a sometimes troubling perception of the duality of virtual agents with their geophysical personae.

A. COMPLEXITY

The officers surveyed believed that the uncertainty they had experienced in responding to cyber-attacks resulted in part from the complexity of the response processes they used, and that understanding the response processes required "an in-depth mastery of cyber warfare tactics, techniques and procedures."[13] This complexity led to ambiguity in lines of authority and actionable thresholds of adversary activity, or "red-lines,"[14] in determining a proper response.[15] This internal "fog of war" in the decision-making process was exacerbated by the lack of scalable response options,[16] which the officers believed limited the ability to respond to the wide range of cyber-attacks.[17]

---

[10] Daryl L. Caudle, Decision-Making Uncertainty and the Use of Force in Cyberspace: A Phenomenological Study of Military Officers, 221, DTIC X (Oct. 14, 2010) (Ph.D. dissertation, University of Phoenix), *available at* www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA534888. The author is grateful for CAPT Caudle's insightful comments and suggestions on this paper.

[11] *Id.* at 225.

[12] *Id.* at 226.

[13] *Id.* at 253.

[14] Accordingly, cyber response decisions are complicated by the assessment of tradeoffs between operational gain and intelligence loss. *Id.* at 261. *See* Ellen Nakashima, *Dismantling of Saudi-CIA Web site illustrates need for clearer cyberwar policies*, washingtonpost.com (Mar. 19, 201), *available at* http://www.washingtonpost.com/wp-dyn/content/article/2010/03/18/AR2010031805464_pf.html (conflict between U.S. Army and the CIA on whether to take down a joint Saudi intelligence-CIA web site used to collect information on potential jihadists, but also used by jihadists to plan operations against U.S. Army units).

[15] Study participants assessed that the planning and conduct of cyber operations is hampered by unwieldy targeting processes and competing interagency interests, a lack of transparency among other government agencies, and unnecessarily classifying information at too high a level compound these problems. Caudle, *supra* note 10, at 255, 261, 263.

[16] *Id.* at 253.

[17] *Id.* at 254.

The study officers also noted the negative impact of an external aspect of complexity in cyberspace; chaos. Chaos describes the phenomenon observed in dynamic, complex systems in which small variations among initial inputs into the systems lead to large variations among the results of these inputs, often in a seemingly random manner.[18] These systems are deterministic, however, and "normally achieve equilibrium around a confined region of space, called a *strange attractor*, where the system permanently resides,"[19] i.e., the precise result cannot be predicted, but it will be of a reasonably foreseeable nature. In practical terms, the impact of chaos on the battlefield has long been recognised.[20] For a cyber commander, chaos means that the same response actions within very similar operational contexts will not always have the same effects, and this uncertainty as to unintended effects further complicates decision-making.[21]

B. TECHNICAL CHALLENGES

"Because cyber warfare is an emerging, non-kinetic, asymmetric warfighting discipline that occurs in a virtual domain, leaders lack experience; moreover, physical effects from the actions they take are not always observable."[22] This likely makes it more difficult in advance of an attack for a responsible cyber commander to make an accurate assessment of the battle-space which would be necessary to support a proper analysis of proportionality and military necessity.[23] Likewise, battle damage assessment is more complex than in the geophysical world,[24] and this likely has a feedback effect on cyber commanders' ability to learn from their experiences in terms of making future assessments of proportionality and military necessity.

Creating effective and integrated hardware and software that would allow a cyber commander to visualize the area of cyber operations accurately will likely require the automatic analysis of multiple sources of data, and the combination of analysis

---

[18] *Id.* at 131.

[19] *Id.*

[20] Alan Beyerchen, *Clausewitz, Nonlinearity and the Unpredictability of War*, 17 Int'l Sec. 59, 70 (1992).

[21] *See* Simon R. Atkinson & James Moffat, The Agile Organization: From Informal Networks To Complex Effects And Agility, 77-84 (2005) (Networked decision-making by Allied forces in the Battle of the Atlantic more capable of dealing with intricacy and uncertainty than German forces), *available at* http://www.au.af.mil/au/awc/awcgate/ccrp/atkinson_agile.pdf.

[22] Caudle, *supra* note 10, at 76.

[23] Neil C. Rowe, *The Ethics of Cyberweapons in Warfare* (2009), *available at* http://faculty.nps.edu/ncrowe/ethics_of_cyberweapons_09.htm (last visited Dec. 29, 2012).

[24] *Id.* Participants in the study believed that "undefined information valuation standards (i.e., clear, consistent, and generally accepted expression of the worth of information)" impeded the conduct of battle damage assessment after a cyber-attack. Caudle, *supra* note 10, at 256.

and data from these multiple and diverse sources.[25] These systems would also need to be able to manage the sensors that collect information about the cyber battle-space, and must be essentially defect-free in the performance of their gathering, managing, and analysis functions. Further, the integrity of the data upon which the systems rely must be defended. Such reliability would allow cyber commanders to trust these systems in making their own decisions, rather than second-guessing the systems.[26]

The surveyed officers were also very concerned with the technical challenges they encountered in discriminating between those things that appeared to be valid military targets and those which were protected because of their civilian nature.[27] Identifying a cyber attacker both accurately and quickly enough to allow for an effective response might be one of the most ambiguous and difficult hurdles to overcome in waging LOAC-compliant cyber warfare.[28] Additionally, difficulty in discerning patterns in attacks as to source and potential severity increased the cyber officers' concerns as to causing unintended, higher level effects, and led to their decision-making cycles being prolonged.[29] More pointedly, the officers believed "the level of attribution certainty required to respond to a cyber-attack [was] arbitrary and unrealistically high."[30]

C. LOAC AND ROE APPLICATION

The study participants acknowledged the applicability of treaties, laws and policy directives to decision-making following a cyber-attack, but found these authorities contributed to the uncertainty in formulating the appropriate response.[31] In particular, the officers "found applying the conventional rules of warfare in

---

[25] Computational Methods for Decision-making, Special Notice 12-SN-0009, Special Program Announcement for 2012 Office of Naval Research, 1-2 (2012) (request for proposals), *available at* http://www.onr.navy.mil/~/media/Files/Funding-Announcements/Special-Notice/2012/12-SN-0009.ashx (last visited Dec. 28, 2012) [hereinafter "Computational Methods"]. General Alexander, the NSA Director and U.S. Cyber Command commander, has stated that it is difficult to obtain a common operational picture of the relevant portions of cyberspace in real time to support operations. Ford, *supra* note 9.

[26] *Id*. at 1-2. Conflicts in equities between agencies regarding cyber actions may be compounded by the difficulty in having stakeholders able to "visualize cyber war and cyber damages." *See* Caudle, *supra* note 10, at 256

[27] Caudle, *supra* note 10, at 254.

[28] Matthew M. Hurley, *For and from Cyberspace: Conceptualizing Cyber Intelligence, Surveillance, and Reconnaissance*, 26 Air & Space Power J. 12, 19 (2012).

[29] Caudle, *supra* note 10, at 254. Surveyed officers noted that decision-making uncertainty "is strongly influenced by the 'fog of war' created in cyberspace resulting from advanced deception capabilities and methods." *Id.* at 255.

[30] *Id.* at 260.

[31] *Id.*

cyberspace to be challenging and not straight forward."[32] Although they appeared familiar with literature suggesting that extant LOAC was applicable when "cyber attack [could] be represented by equivalent kinetic attack characteristics," the officers did not believe the current legal framework addressed "sovereignty challenges, jurisdiction boundary problems (e.g., cloud computing and transborder data flows), non-state actors, severity thresholds, and the technical nuances of cyber attack."[33] The officers noted significant concerns with the ROE under which they had operated, which they assessed as "nascent and generally untested,"[34] and in particular they found that "the lack of practical definitions for hostile intent and hostile act in cyberspace" made consistent cyber responses difficult.[35]

D. VIRTUAL AND GEOPHYSICAL DUALITY

For purposes of this article, perhaps one of the most interesting findings of this study was the cyber warfare officers' perception of the "complex duality of [cyberspace's] virtual and physical nature."[36] This both complicated the understanding of higher-order effects in decision-making, and the deconfliction of "traditional military activities from intelligence gathering activities."[37] This duality is further reflected in their perception of their adversary human counterparts being virtually coupled with their digital agents.[38] To improve decision-making, the respondents believed that policies and ROE that dealt with cyber actions conducted solely within cyberspace were necessary to "depersonalise" cyber-attacks, and that this depersonalisation would reduce uncertainty and make cyber responses quicker.[39]

E. SUMMARY

The scope of the challenges identified by the cyber warfare officers illustrates just how very different cyber conflict is from geophysical conflict, as well as the clash between these differences and the perfectly human desire to understand cyber conflict in terms consistent with, or at least analogous to, geophysical human

---

[32] *Id.* at 255. Study participants described the current legal framework as "antiquated and inadequate to support military operations in an effective manner." *Id.* at 261.

[33] *Id.* at 262.

[34] *Id.* at 260.

[35] *Id.* at 262.

[36] *Id.* at 266. As one writer has noted, "[i]n the virtual world, when we refer to an enemy or an opponent, we may actually be referring to what really are the second and third order effects of the actual activity of our opponent, or even beyond." The Basics of Cyber Warfare: Understanding the Fundamentals in Theory and Practice, 67 (Steve Winterfield & Jason Andress, eds.,2012).

[37] Caudle, *supra* note 10, at 266.

[38] *Id.* at 256.

[39] *Id.* at 258.

experiences. This suggests that there will likely be a heavy burden upon system designers to accommodate the human aspect of decision-making in the development of hardware and software intended to support cyber commanders and operators. In particular, the officers' concerns as to the perceived duality of cyberspace actors suggests that proposed solutions must be mindful of the need to deliberately and explicitly differentiate between cyber agents and their geophysical personae when it would increase operational efficiency, and to foster this duality when it would have the same effect.

3. BUYING TIME: ALLOWING THE COMMANDER TO BE RESPONSIBLE

Given the concerns detailed by the surveyed officers as to the conduct of cyber operations, and the misgivings of many regarding the use of autonomous systems waging war, it is useful to consider alternative methods of compressing the temporal aspect of cyber decision cycles while ensuring the responsible cyber commander remains in the decision loop. These methods include improved cyber intelligence, surveillance and reconnaissance (ISR) capabilities, innovative staffing techniques, enhanced human-computer interfaces (HCIs), and possibly even brain-computer interfaces (BCIs).

A. *CYBER ISR*

In light of the current emphasis on defensive and offensive cyber operations in the U.S. national and military doctrine and policy, one writer has assessed the ISR aspect of cyber operations as particularly needing "doctrinal, educational, and organizational concepts that forcefully emphasize the centrality and operational nature of cyber ISR."[40] Cyber ISR can take many forms, with different levels of intrusiveness into potential adversaries' cyber infrastructure. At one end of the intrusiveness spectrum, "honeypots" or "honeynets" can be emplaced in a cyber system's defences to lure intruders to penetrate them instead,[41] thereby providing a cyber commander with advanced warning of potential attacks. At the other end of the spectrum is the use of "active defence" mechanisms that operate within an adversary's cyberspace, "exploit[ing] [its] cyberspace vulnerabilities while gaining

---

[40] Hurley, *supra* note 28, at 20-21. Enhanced cyber ISR would likely need to include robust indications and warnings (I &W) intelligence processes to be effective. *See* Chairman, Joint Chiefs of Staff, Joint Publication, 2-0, Joint Intelligence , I-16-17 (Jun. 22, 2007) (I & W intelligence is "very time-sensitive" forewarning of adversary actions or intentions).

[41] *See* Lance Spitzner, Honeypots: Tracking Hackers, 50-71 (2002) (discussing the operational value of honeypots and honeypot networks).

a deeper understanding of the enemy's decision cycle and defensive weaknesses."[42] Theoretically, this too should allow more time for responsible cyber commanders' to make decisions.

Such measures, however, inevitably lead to effective countermeasures.[43] Rather than compressing decision cycles through the provision of better informational inputs, the continuous ISR race between cyber adversaries might in the end only result in maintenance of the decision cycle status quo. Further, the use of active defence measures could result in a cyber adversary interpreting such probing as an indication of hostile intent, or a hostile act,[44] and decide to engage in an unexpected, forceful counter-response it might otherwise not have conducted.

B. STAFFING TECHNIQUES

A second approach would be the acceleration of cyber commanders' decision-making processes through innovative staffing techniques. This could include the use of multiple planners and multiple commanders with cyber weapons release authority working within a cyber operations centre, each supported by teams of technical advisers (TEKADs), political advisers (POLADs), and legal advisers (LEGADs).[45] The use of teams of commanders would allow multiple emergent situations to be dealt with simultaneously. Although weapons release authorities and processes are not currently structured this way for kinetic operations, multiple commanders could possibly be tiered, so that increasing levels of likely incidental cyber or geophysical damage or injury could be handled by progressively senior commanders. The commanders' reaction times could likely be reduced if they were supported by dedicated adviser teams with whom they habitually trained and operated, particularly if they were using clear ROE.[46] Unfortunately, whilst such staffing measures could possibly reduce a cyber commander's decision cycle by minutes, the operational flow within cyberspace might be moving too fast for such a reduction to make a meaningful difference.

---

[42] Caudle, *supra* note 10, at 77.

[43] *See, e.g.*, Neil C. Rowe & Han C. Goh, *Thwarting Cyber-Attack Reconnaissance with Inconsistency and Deception*, Proceedings of the 2007 IEEE Workshop on Information Assurance, 151, 151-58 (U.S. Military Academy, West Point, NY, June 20-22, 2007).

[44] *See* Timothy L. Thomas, *China's Electronic Long-Range Reconnaissance*, 87 Military Review 47, 47-54 (Nov./Dec. 2007) (discussing suspected Chinese cyber intrusions in the context of Chinese cyber strategy and reconnaissance's role in the Chinese concept of "active offense").

[45] *See* Tallinn Manual, *supra* note 1, Rule 52, para. 6, 158 ("mission planners should, where feasible, have technical experts available to assist them in determining whether appropriate precautionary measures have been taken.").

[46] General Alexander has suggested that clear ROE might speed up cyber decision-making. Ford, *supra* note 9.

C. HCIs

A third approach, heightening the intimacy of the connections between cyber commanders and their computer systems,[47] could occur in two primary ways: enhanced human-computer interfaces (HCIs) and brain-computer interfaces (BCIs). As to HCIs, the software that creates the operational picture for the commander might be tailored to that specific commander's personality traits. Research has shown that students using computer interfaces that recognized their individual learning styles showed higher learning results.[48] Further, technologies such as deep-learning programmes, which use artificial neural nets to imitate the way the human brain learns, offer the promise of computers that could both recognize patterns in large amounts of information and then communicate this to humans via speech.[49] Such programmes have already displayed what appears to be the capability to learn as they recognize patterns.[50]

The use of HCIs that allow interactions similar to how human-human interactions occur, so that computers could potentially understand and or anticipate human intentions,[51] would conceivably allow for a commander to react more quickly to emergent situations in cyberspace. Through enhanced communication, these systems could shorten cyber commanders' decision cycles by presenting cyberspace visualisations specifically attuned to particular commanders' problem-solving and interaction styles. As with staffing innovations, however, the decreases in time might simply not add any operational advantage.

D. BCIs

At one level, given the physical invasiveness of some BCI techniques, such connections resemble science-fiction, but recent advances in this field have been remarkable. For example, a user has been able to move an automated prosthetic arm and grasp items simply through thought, as her brain activity was registered

---

47 Adams, *supra* note 8, at 66.

48 Edmond Abrahamian, Jerry Weinberg, Michael Grady, and C. Michael Stanton, *Is Learning Enhanced by Personality-Aware Computer-Human Interfaces?,* Proceedings of I-KNOW '03, 226, 228-29 (Graz, Austria, July 2-4, 2003).

49 John Markoff, *Scientists See Promise in Deep-Learning Programs*, NYTimes.com (Nov. 23, 2012), *available at* http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?_r=0.

50 John Markoff, *How Many Computers to Identify a Cat?* 16,000, NYTimes.com (June 25, 2012), *available at* http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all.

51 Hayrettin Gürkök & Anton Nijholt, *Brain Computer Interfaces for Multimodal Interaction: A Survey and Principles*, 28 Int'l J. Human-Computer Interaction 292, 292-93 (2012).

by microelectrodes implanted in her brain.[52] Certain techniques do not require physical contact between users' brains and the computer systems, but instead monitor brain activity through contact sensors on the users' scalps.[53] The potential melding of multimodal interaction techniques, in which computers use multiple sensors to gather physical information about interacting human partners (e.g., cameras to watch hand movements and eye gaze, microphones to register spoken commands) with BCIs that directly track human brain activity[54] offers a possible future means to further speed up a commander's decision-making.

On balance, however, ethical and technological challenges suggest that this approach is likely to be of limited use in cyber conflict in the near term.[55] Further, invasive interfaces would appear to raise the possibility of directly targeting the human operator through cyber attack. Creating a vulnerability that allows the specific targeting of a highly trained commander makes little operational sense.

E. SUMMARY

Use of these different approaches, possibly in combination, could yield significant improvements in the response times of cyber commanders while ensuring their actions remain LOAC-compliant. Currently, however, it does not appear that any single approach or combination of approaches would satisfy the operational imperative to be able to respond to emergent cyber situations as quickly as ADPs could. Therefore, it is important to explore how ADPs might be constructed and employed in cyber operations to provide commanders means of effective engagement.

4. BUILDING LOAC-COMPLIANT ADPs

The increasing use of military unmanned aerial vehicles (UAVs) operated by human controllers has triggered constructive discussion regarding the ethics, legality, and practicality of such weapon systems becoming autonomous.[56] The issues raised

---

[52] Jennifer L. Collinger et al, *High-performance neuroprosthetic control by an individual with tetraplegia*, thelancet.com, 6-7 (Dec. 17. 2012), *available at* http://dx.doi.org/10/1016/S0140-6736(12)61816-9.

[53] Alessandro Pressaco et al, *Neural decoding of treadmill walking from non-invasive, electroencephalographic (EEG) signals*, J. Neurophysiology, 5-6 (July 13, 2011), *available at* http://jn.physiology.org/content/early/2011/07/11/jn.00104.2011.full.pdf+html.

[54] *See* Gürkok, *supra* note 51, at 303-04.

[55] *See* Jens Clausen, *Moving minds: Ethical Aspects of neural motor prostheses*, 3 Biotechnology Journal 1493, 1496-98 (2008), *available at* http://www.yorku.ca/lsergio/Clausen_MovingMindsBTJ2008.pdf (medical complications, interference with personality and personal identity, and responsibility for malfunctions are among the ethical issues raised by BCI).

[56] Arkin, *supra* note 7, at 21-25.

therein are directly applicable to the proper relationship between cyber conflict ADPs and human commanders. Consideration of the advantages and disadvantages of using such systems, the potential architecture of LOAC-compliant ADPs, and the human-robot interaction (HRI) aspect of the operation of these ADPs all suggest that the role of the responsible cyber commander could be appropriately factored into their design and use.

A. *POSSIBLE ADVANTAGES OF ADPs*

ADP proponents argue that these systems could enhance the observance and application of LOAC and complementary ROE in conflicts.[57] First, human combatants endure physical pressures in conflict that degrade human perception and the rational decision-making based upon it.[58] These stressors generate negative emotions that further degrade both perception and cognition.[59] ADPs would not be subject to these physical stressors, and the programmed responses would not be clouded by emotion.[60] Second, because of their potential to receive and integrate vast amounts of information quickly from multiple sensor systems, the decisions made might be based on a more complete picture of the cyber operational area than a human could comprehend.[61] Further, this operational picture could be evaluated without "the human psychological problem of 'scenario fulfilment'" in which "humans use new incoming information in ways that only fit their pre-existing belief patterns.'"[62] Third, ADPs could serve as independent witnesses of the cyber action, whose decisions to record and report potential violations of LOAC and ROE are not subject to concerns of disloyalty to fellow soldiers.[63] Fourth, one human operator might be capable of simultaneously overseeing multiple ADPs.[64]

---

[57] *Id.* at xv-xviii.

[58] Helmet-Mounted Displays: Sensation, Perception and Cognition Issues, 675-749 (Clarence E. Rash et al, eds., 2009).

[59] *See* Arkin, *supra* note 7, at 33-36 (robots would not engage in irrational thinking that tends to dehumanize adversaries and excuse their lethal engagement on the basis of genocidal, penal, or utilitarian rationales).

[60] Ministry of Defence, Joint Doctrine Note 2/11, The UK Approach to Unmanned Aircraft Systems, 5-11 (Mar. 30, 2011), *available at* http://www.mod.uk/NR/rdonlyres/F9335CB2-73FC-4761-A428-DB7DF4BEC02C/0/20110505JDN_211_UAS_v2U.pdf. [hereinafter "UK Approach"].

[61] Arkin, *supra* note 7, at 29-30; Gary E. Merchant et al, *International Governance of Autonomous Military Robots*, 12 Colum. Sci. & Tech. L. Rev. 272, 279-280 (2011).

[62] Arkin, *supra* note 7, at 29-30; Merchant, *supra* note 61, at 279-80.

[63] *Id.*

[64] UK Approach, *supra* note 60, at 5-10.

B. *POSSIBLE DISADVANTAGES OF ADPs*

Critics of ADPs believe that their use would actually result in fewer LOAC-compliant decisions. First, they suggest that human emotions are actually a safeguard against killing,[65] because currently only humans have the ability to "bring empathy and morality to complex decision-making,"[66] and that the human ability to factor emotion into assessments of hostile intent is crucial when decision-makers are dealing with human behaviour.[67] Second, human operators might experience "automation bias," and be unwilling to challenge an ADP's assessment or action.[68] Third, ADPs cannot be made sophisticated enough to make the context-dependent assessments that human commanders make on the basis of incomplete information, such as proportionality.[69] Similarly, ADPs will not have sufficient capability to distinguish between civilians and combatants.[70] Even assuming sufficiently sophisticated software could be developed to allow ADPs to make such decisions, perhaps using artificial intelligence,[71] "[c]omputer programs do not behave as predictably as software programmers would hope."[72] Complex systems are subject to malfunctions, and '[p]ortions of programs may interact in unexpected, untested ways."[73] Complexity itself might generate non-programmed and unanticipated emergent behaviours.[74] Even an ability to learn, which would appear desirable from the viewpoint of creating a system that could adapt to novel situations, "raises the question of whether it can be predicted with reasonable certainty *what* the [ADP] will learn."[75] Further, critics believe the use of ADPs will lead to a "responsibility gap," because there is no fair and effective way to hold humans responsible for the effects of automated decision-making when they had no direct control over the decision-making process.[76]

---

[65] Human Rights Watch & International Human Rights Clinic, Losing Humanity: The Case against Killer Robots, 37 (2012) [hereinafter "Losing Humanity"].

[66] *See* UK Approach, *supra* note 60, at 5-11

[67] Losing Humanity, *supra* note 65, at 31-32; Merchant, *supra* note 61, at 283.

[68] Losing Humanity, *supra* note 65, at 13.

[69] *Id.* at 42; Merchant, *supra* note 61, at 285.

[70] Losing Humanity, *supra* note 65, at 20.

[71] UK Approach, *supra* note 60, at 5-4. "[S]uch operations would present a considerable technological challenge and the software testing and certification for such a system would be extremely expensive and time consuming."

[72] Merchant, *supra* note 61, at 284.

[73] *Id.* at 283-284.

[74] *Id.* at 284.

[75] *Id.* (emphasis in original). This is particularly of concern if the robot operates in an unstructured environment.

[76] Losing Humanity, *supra* note 65, at 42.

C. ARCHITECTURE

Professor Arkin has posited that there are three primary requirements that must be met for an ADP to respond to a situation in conformance with ethical parameters, such as LOAC and ROE. These are the ability to perceive the operational environment correctly, the inclusion of content which identifies the specific types of acts permitted or prohibited under these parameters, and the appropriate representation of this content within the decision architecture.[77] Arkin advocates programming which essentially errs on the side of caution so that the context and nuance upon which a human commander would rely becomes non-relevant in the ADP. For example, a certain continuing level of quality in the situational awareness upon which the ADP relies could be required before it could respond. Requiring this threshold to be met would enhance target discrimination, and prevent the engagement of civilians, civilian objects or friendly forces throughout the course of an engagement.[78] One component of this threshold could be the requirement to have consistent information from multiple sensors.[79]

As to possible ADP responses, Arkin suggests first that the decision architecture should be designed so that ethical responsibility is segregated within it.[80] Arkin sees four separate functions as being necessary to ensure conformance with ethical standards, the first of which would be an "ethical governor," which would "conduct an evaluation of the ethical appropriateness" of any ADP-proposed lethal response.[81] The ethical governor would be complemented by "ethical behavior controls," which would only allow the system to propose responses consistent with LOAC and ROE,[82] and by "ethical adaptors," which would monitor on-going cyber responses and essentially call "cease fire" if certain thresholds were exceeded.[83] With UAVs in the geophysical world, this requirement is satisfied by means of near real-time video feeds that provide commanders and their advisers with an understandable operational picture of the target site.[84] The fourth component would be a "responsibility advisor," which would be "part of the human-[computer]

---

[77] Arkin, *supra* note 7, at 70-91.

[78] *Id.* at 119.

[79] *Id.* at 120-21.

[80] *Id*. at 126.

[81] *Id.* at 127. The governor essentially serves as a cross-check on the response proposed by the ADP. *Id.* at 125.

[82] *Id.* at 133. Further, actions deemed to violate LOAC requirements as programmed would never be undertaken, nor would actions permitted under ROE but in violation of LOAC. *Id.* at 212.

[83] *Id.* at 138.

[84] *See* UK Approach, *supra* note 60, at 5-1, n.2 (UAVs in Afghanistan use the same ROE and targeting guidance as manned aircraft, "but they have the persistence to check and re-check, possibly via legal advisers, that they are compliant" with the ROE).

interaction component" that is used to secure permission from the commander for the ADP to engage in the mission, and to allow commander overrides of ADP decisions.[85]

From an engineering perspective, the legal framework for operating the ADP could essentially be treated the same as other technical and operating requirements at the beginning of the design, so that it could be referenced in the "specification and design of various subsystems, as well as informing the concept of employment."[86] There are different models for ethical decision-making in the context of LOAC and ROE that could be utilised, and this suggests that LEGADs should be part of the software development team.[87] Legal review of the information that would be considered by the ADP would likewise be required, so that the achievable level of situational awareness can be understood[88] in the context of LOAC compliance.

D. HRI

HRI is a relatively new field that addresses in a multidisciplinary manner how people work or play with robots rather than computers or tools, and "[t]his large multi-disciplinary mix presents a very different mindset from traditional engineering design, interface development or ergonomics."[89] Research into HRI so far largely replicates findings from human-human research, and interestingly, shows that "humans expect unmanned systems to meet expectations of a team member with known competences."[90] This is perhaps in its own way a reflection of the cyber/geophysical duality that the surveyed cyber officers noted in their perceptions of cyberspace actors. To meet these expectations, "[m]odels of what operators or decision-makers need to know about the system or state in order to maintain trust in the predictable outcomes from using the system"[91] would need to be developed.

Perhaps because HRI is so new, it is not clear that designers of ADPs fully recognize how important human-friendly perception of the battle-space in which the ADPs operate is for humans to be able to work effectively with the ADPs.[92] As one report

---

85 Arkin, *supra* note 7, at 143.

86 UK Approach, *supra* note 60, at 5-2.

87 Arkin, *supra* note 7, at 95-113.

88 *See* UK Approach, *supra* note 60, at 5-3.

89 Defense Science Board, Department of Defense, Task Force Report: The Role of Autonomy in DoD Systems, 44 (July 2012), *available at* http://www.acq.osd.mil/dsb/reports/AutonomyReport.pdf [hereinafter "Role of Autonomy"].

90 *Id.* at 46.

91 *Id.* at 49.

92 *See id.* at 23 ("For the operator, autonomy is experienced as human-machine collaboration, which is often overlooked in design.").

assessing current autonomous military systems noted, this crucial aspect "is largely ignored and instead erroneously treated as a computer display problem; however, a display cannot compensate for a lack of sensing."[93] Further, "[m]ultisensor integration, either for increased sensing certainty or more comprehensive world modelling, appears to be ignored."[94] As a result, areas of deficiency in human-system collaboration include the lack of "natural user interfaces enabling trusted human-system collaboration and understandable autonomous system behaviors."[95] This could be remedied through the creation of "perceptually oriented interfaces and sensor placement designed around the psycho-physical attributes of the human perceptual system," such as enabling dialogue between humans and ADPs "using natural human interaction modes, especially natural language and gestures."[96]

Not properly addressing the need for optimal human interface in design architecture causes commanders and operators to lack confidence that the systems will operate as they are supposed to, and this lack of trust[97] in turn likely limits the systems' usefulness and the speed at which decisions can be made.[98] This gap in confidence could possibly be remedied by an emphasis on "natural user interfaces and trusted human-system collaboration, perception and situational awareness to operate in a complex battle-space, large-scale teaming of manned and unmanned systems, and test and evaluation of autonomous systems."[99] These improvements would provide the human partner sufficient visibility of the system's activities and how they related to the mission objectives,[100] and would also likely help normalise the legal review process as well.[101]

Concerns of ADP critics to the contrary, the proper assignment of responsibility for the use of ADPs is likely easily resolved – it will lie "with the last person to issue the command authorising a specific activity."[102] Such authorisations can be reliably recorded in a log for audit trail purposes; for example, as part of the preparation process for a cyber commander assuming a watch in an operations centre.[103] In fairness, however, for the authorising commander to be held responsible there would

---

93 *Id.* at 36.

94 *Id.* at 37.

95 *Id.* at 48.

96 *Id.*

97 *Id.* at 2.

98 *See id.* at 1 (misperceptions as to the meaning and implications of autonomy are limiting its adoption in the military).

99 *Id.* at 8-9.

100 *Id.* at 48.

101 UK Approach, *supra* note 60, at 5-3.

102 *Id.* at 5-5.

103 *Id.* at 5-6.

need to be an underlying "assumption that a system will continue to behave in a predictable manner after commands are issued."[104] Reliance upon this assumption would become more problematic "as systems become more complex and operate for extended periods" without human intervention.[105] Fostering the level of trust in the operation of ADPs necessary for a commander to decide affirmatively to be responsible for their effects would likely require an extensive, holistic development of complementary education, realistic training, and user-friendly ADP hardware and software.[106]

E. SUMMARY

At this point in time, it does not appear to be technologically feasible, nor is it necessary, for an ADP to attempt to quantify the moral and emotional differences between responses to a potential targeting problem. The practical effect of Professor Arkin's proposal for programming ADPs is to have them unable to make the close calls, and therefore unable to engage on their own unless quantifiable criteria which have been established by erring on the side of caution have been met. The close calls, and the potential use of emotion and morality, are reserved for the human member of the team.

5. CONCLUSION

Ensuring that commanders remain responsible in the course of cyber conflict will first require significant investment in the sensors, machines, and software that are used to provide the operational visualisation of cyberspace and the geophysical world upon which the commanders would rely when making use of force decisions. This would include the ability to map the relevant portion of cyberspace to identify those points at which cyber action might reasonably be expected to ripple into the geophysical world. Commanders must be confident that the situational information and the analysis derived from these systems are accurate, and they must have the ability to access geophysical surveillance and reconnaissance assets that could provide them near real-time awareness of the likely ripple points.

Second, ADPs must be developed that quickly assess this situational information and analysis in a conservative fashion so that commanders are alerted when proposed cyber responses could be reasonably expected to cause injuries to humans or damage

---

[104] *Id.* at 5-5.

[105] *Id.* at 5-5.

[106] *See* Caudle, *supra* note 10, at 259, 273, 280 (new doctrine and training required to optimise cyber commander performance).

to tangible objects; or violate pre-set red-lines in terms of significant damage to data in targeted systems. Such systems might employ expanded consideration of indirect effects to reassure commanders that they are being provided a satisfactorily holistic assessment, and as a preventive measure to keep either geophysical or cyberspace thresholds from being crossed automatically or inadvertently. This would give cyber commanders confidence that their decisions to engage in cyber actions were not made in too narrow a fashion.

Third, if instances of human injury or damage to geophysical objects could be reasonably expected, then a commander at some level, rather than an ADP, would need to make the affirmative decision to engage after satisfying LOAC and ROE requirements. If no geophysical injury or damage was reasonably expected, but red-lines regarding cyber infrastructure were reasonably likely to be approached within a certain margin of uncertainty, a cyber commander at some level would also need to decide whether and how to respond within the prescribed ROE, which might themselves contain LOAC-like decision factors. If the effects of the proposed cyber action would occur and remain solely within cyberspace, then ADPs could conceivably operate without human intervention using default settings based on approved ROE. For a commander to remain confident that ADPs were behaving as expected, though, continued monitoring of the cyber and geophysical aspects of the battle-space would be necessary to ensure human override thresholds were not reached until the cyber action was completed.

Presumably, since the vast bulk of cyber action would occur within cyberspace and the effects would remain there, keeping cyber commanders responsible and compliant with LOAC should be achievable. Proposed cyber actions that could be reasonably expected to ripple into the geophysical world and cause human injury or damage to objects might likely prove to be the exception rather than the rule. Cyber actions that might violate cyberspace ROE red-lines regarding cyber infrastructure would likely be more common, but these activities would not entail potential criminal violations of international law at this point, and their effects might be both reversible and quickly terminated. For ADPs to be used properly, however, military organisations must begin investing in the education and training curricula and opportunities that that will be required to groom young cyber operators for their future roles as effective and responsible cyber commanders.[107]

---

[107] *Id.* at 279-80.