Lethal Autonomous Weapons: Meaningful human control or meaningful human certification?

M.L. Cummings

There has been increasing debate over the use of autonomous weapons in the military, whether they should be banned for offensive uses, and even whether such technologies threaten our very existence[1]. As both a former fighter pilot for the US Navy but also a professor of robotics, I find these debates filled with a lack of technical literacy and emotional rhetoric, often made worse by media and activist organizations that use fear to drive exposure and funding. As I will discuss in more detail, the technology that enables autonomous behavior is nowhere near ready for use on the battlefield, and caution is needed. However, autonomous technologies may be able to fill a critical gap in human decision making on today's battlefield that often leads to unnecessary accidents and deaths. Given my unique position on both sides of the technology development and operational coin, I offer up the following points for consideration in this very complicated and multi-dimensional issue.

Autonomous vs. Automated

First, it is important to define autonomy in technology, which is not the same as automation. Automated systems operate by clear repeatable rules based on unambiguous sensed data. Autonomous systems take in data about the unstructured world around them, process that data to generate information, and generate alternatives and make decisions in the face of uncertainty. Often, people will refer to this set of capabilities as self-governing. Systems are not necessarily either fully automated or fully autonomous, but often fall somewhere in between.

Automated defensive weapons systems exist today like the Patriot missile system or Israel's Iron Dome. It has long been recognized that humans simply cannot often respond in time to threats like incoming rocket attacks since they have an inherent limitation known as the neuromuscular lag (Jagacinski & Flach, 2003). Even if a person is paying attention perfectly to an evolving situation, there is still an approximate half second lag for a person to see a problem and then act accordingly. This is why automated defensive systems are given the role and ability to authorize weapons launch.

Automated offensive weapons also exist and are routinely used in Unmanned Aerial Vehicle (UAV, aka drone) missions. One of my jobs as an F/A-18 pilot was to get a radar-guided missile as close as possible to an airborne target (like an enemy helicopter) without being attacked. Once I was close enough to give a missile an initial guess as to the target's location, I would shoot the missile, and then once the missile had a radar lock on the target, run away to get out of the enemy's missile envelope. Drones like the Predator and Reaper do this exact same job, but they leave the human out of the cockpit, and supervising from a safe distance. However, the fundamental roles of the humans and computers do not change in drone missions. Thus, in

[1] For a review of current political, ethical, security and legal implications surrounding these weapons, see (Bibi, 2018) and (Wagner, 2014).

current operations, a human guides a drone to the target, identifies the target, and approves weapons launch.

In contrast to automated weapons, autonomous weapons are given a goal (e.g., find and kill the enemy) and then should execute the missions by themselves until the goal is accomplished. Currently there are no offensive autonomous weapons used in dynamic battlefield scenarios, but as will be discussed in a later section, the US has been using a primitive autonomous weapon, the Tomahawk, for 30 years to strike static targets.

What does meaningful human control actually mean?

Proponents for a ban on offensive autonomous weapons advocate that any use of such weapons should not be beyond meaningful human control (Future of Life Institute, 2015). This language is problematic at best because there are widely varying interpretations of what meaningful human control is. For example, meaningful human control (MHC) could mean that a human has to initiate the launch of such a weapon. In today's networked world, this person can (and often is) thousands of miles from the intended target. Is it truly meaningful for a human to release a weapon if he or she is sitting 4000 miles from the point of weapon release? Does MHC mean a human has to monitor the weapon until impact and perhaps have the ability, however remotely, to abort the mission? Are we asking humans to take on MHC tasks that have high likelihoods of human error and what are the consequences?

Fundamentally this question of meaningful control aligns with a question faced every day by controls engineers who must determine which functions and roles automation/autonomy should have versus what should be allocated to humans when designing a complex system. To answer such questions, engineers and designers must know the strengths and limitations of humans as well as of the technologies, and such boundaries are not always clear. It is often not clear whether a technology or a human should or could be in control from a design standpoint, such as whether self-driving cars are actually better than humans across a number of driving scenarios (Cummings, 2019). Such design gaps inevitably lead to discussions of ethical and social impact of the technology, especially in safety critical systems like transportation and medical systems, as well as autonomous weapons.

To shed more light on these issues, two vignettes of actual military engagements are provided to frame the MHC debate, which is really a discussion about role allocation between humans and autonomous systems. I suggest that what is needed is not a call for meaningful human control of autonomous weapons but rather a focus on meaningful human certification of such systems.

The Chinese Embassy Bombing

During NATO's Operations Allied Force campaign in Yugoslavia in May of 1999, the Chinese embassy in Belgrade was accidentally bombed with a US Air Force GPS-guided Joint Direct Attack Munition (JDAM). Three Chinese journalists were killed, with more than 20 people injured. The original target was a Yugoslav supply warehouse, about 500 yards from the

embassy. The incorrect coordinates were provided to the Air Force by the US Central Intelligence Agency, in part due to faulty maps and a flawed review process that failed to catch the error (Rasmussen, 2007).

Operation Provide Comfort

In 1994, the United States sought to stop Iraqi attacks on Kurds in northern Iraq by enforcing a no-fly zone through Operation Provide Comfort. The no-fly zone was patrolled by US Air Force F-15 fighters, supported by an Airborne Warning and Control System (AWACS) aircraft. On April 14, two US Army Black Hawk helicopters transported U.S., French, British and Turkish commanders, as well as Kurdish personnel across this zone when the F-15 fighters misidentified them as Iraqi Hind helicopters. The fighters launched air-to-air missiles, killing all 26 onboard the Black Hawks. One person on the AWACS plane gave the Black Hawks permission to enter the no-fly zone. However, due to a communication breakdown, this information was not shared with another controller on the same AWACS communicating with the F-15s. One government accident investigation found 130 causal human errors for this single accident (GAO, 1997).

These two case studies exemplify the kinds of missions faced by military operators in deciding when and where to launch an airborne weapon with likely lethal consequences. The Chinese Embassy incident represents what should have been a relatively low uncertainty engagement given the intended target was static and known, while the *Operation Provide Comfort* case was one of high uncertainty with a dynamic and unknown potential threat. Understanding the relative level of uncertainty is key to understanding how much human control could and should be exerted in these scenarios, and what the implications are for the use of autonomous weapons in these settings (Cummings, 2017).

Currently autonomous technologies perform best when they are doing a very narrow task with little to no uncertainty in their environments. For example, driverless cars can follow lines on roads very well as long as there are no long shadows on the road. However, during sunrise and sunset, such systems struggle due to the cameras' inability to cope with solar glare. So, the cars perform under low uncertainty, but terribly under high uncertainty. The following sections describe in more detail the influence of high and low uncertainty scenarios on autonomous weapons and what these issues mean for possible future deployment.

High uncertainty engagements are outside the realm of autonomous weapons – for now

Unfortunately, the problems surrounding the *Operation Provide Comfort* incident are not new to the military and even today are sadly all too common. There are many very similar incidents where a suspected target emerges on or near a battlefield, information about this target is misinterpreted, and because of the "fog of war" which includes incomplete information, psychological biases, and human error, a weapon is launched, killing innocent civilians or friendly forces (Rasmussen, 2007).

One of the biggest problems with high uncertainty targeting situations is time pressure. Identification and classification of such targets is always the highest priority, but when targets

are moving and perceived as a threat, there is limited time for those humans in control (the F-15 pilots in the *Provide Comfort* example) to make a high-quality decision. Research has shown that warfighters under time pressure in the air and on the ground are predisposed to a number of psychological biases that cause them not to evaluate all relevant pieces of information, which in many cases leads to disastrous outcomes (Cummings, 2004; Parasuraman & Manzey, 2010).

Autonomous weapons could, in theory, aid human decision makers in these time-critical targeting scenarios who struggle to gain unbiased information in a timely fashion because autonomous weapons could make decisions quickly and more accurately. While the role of target identification is currently assigned to humans, the U.S. military has spent a significant amount of money over the past 20 or more years to develop various forms of automated target recognition systems for unmapped and dynamic targets, but progress has been slow and difficult. Recent approaches have tried to add more advanced forms of computer vision and machine learning techniques to improve performance, but resulting systems have had high false positive rates and other technical problems (Ratches, 2011).

The limitations of machine learning and computer vision systems are the Achilles' heel for all autonomous systems, military and civilian. Computer vision systems struggle, especially under dynamic and unfamiliar conditions, to integrate incoming sensor data into the "brain" of an autonomous system to form a world model by which action decisions can be made (Cummings, 2017). Despite the hype surrounding the advancement of driverless cars that leverage artificial intelligence in their computer vision systems, they and all such autonomous systems are deeply flawed in their ability to reliably identify objects (Cummings, 2019; Marcus, 2018).

Figure 1 illustrates just how computer vision flaws using a deep learning algorithm are manifested in identifying vehicles in different poses. In all three examples, a typical road vehicle (school bus, motor scooter, firetruck) is shown in a normal pose, with 3 other unusual poses. While a bus on its side may be a rare occurrence in everyday driving, such unusual poses are part of the typical battlefield environment. Deep learning data-driven approaches to reasoning are extremely brittle and often unable to cope with even the smallest perturbation of data.

Given these fundamental problems in current computer vision technology that make it very unreliable in understanding the world, especially in dynamic situations, any weapon system that requires autonomous reasoning based on machine learning, either offensive or defensive, will be deeply flawed. Moreover, it is entirely likely that these computer vision weaknesses could be exploited and create new cybersecurity concerns, which have already been demonstrated in street sign (Evtimov et al., 2017) and face recognition (Sharif, Bhagavatula, Bauer, & Reiter, 2016) applications. Because of these technological limitations, at the current time, the role of acquiring a target is still very much a function for humans to execute.

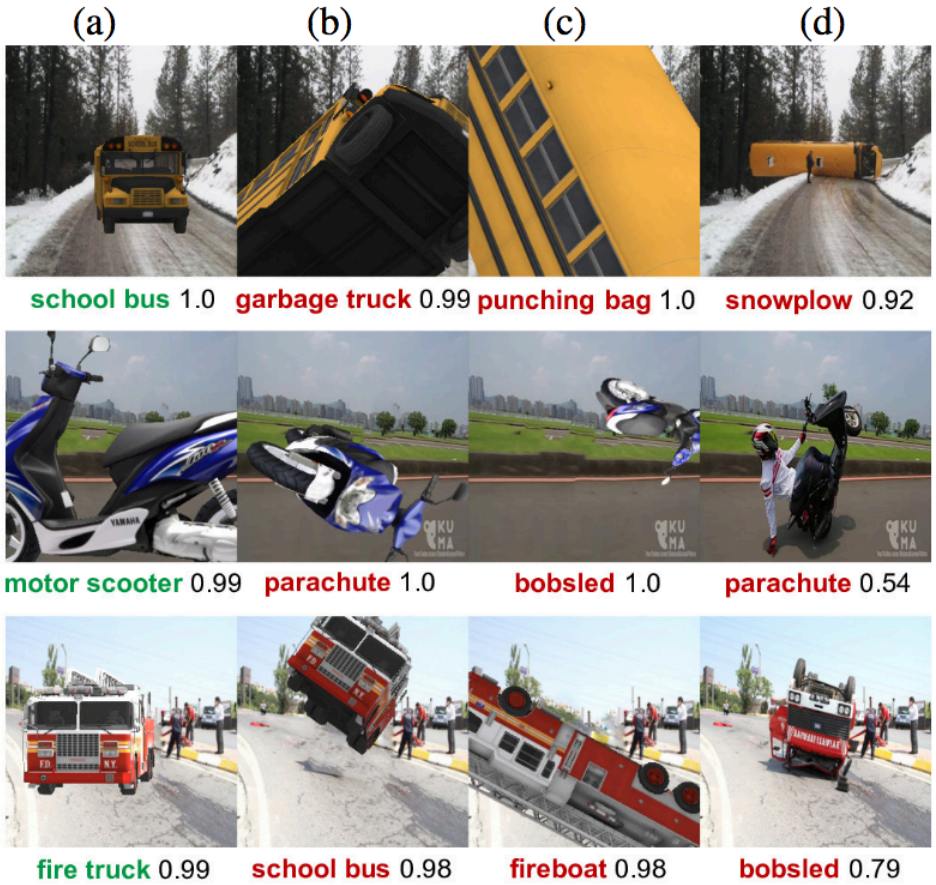

Figure 1: A deep learning algorithm prediction (probabilities following the algorithm's label) for typical road vehicle poses in a 3D simulator (a) and for unusual poses (b-d) (Alcorn et al., 2018).)

From a pure engineering assessment approach, any autonomous system that currently relies on computer vision systems to reason about dynamic environments is likely to be extremely unreliable, especially in situations never before encountered by the system. Unfortunately, this is exactly the nature of warfare. Thus, maintaining human control in current military operations is critical, since the technology is so deeply flawed. But while the technology currently cannot cope with uncertainty on the battlefield, it is far too premature to say that will always be the case. As the *Operation Provide Comfort* example illustrates, while humans are better than current deep learning algorithms at consistent target identification, they are very much error prone in real world scenarios. As will be discussed in a later section, ***if*** autonomous targeting systems in dynamic settings could be shown to be superior to humans at some point in the future, which is a very high bar, I argue that not only should we use them, but we have an obligation to do so.

This previous section focused on the concept of meaningful human control and dynamic targeting, but there are additional lessons to be learned by examining autonomous weapons and static targeting, discussed in the following section.

Low uncertainty engagements are very much in the realm of autonomous weapons today

For dynamic targeting in offensive missions, the previous section outlined that because of serious flaws in computer vision algorithms and high uncertainty, human control is necessary due to the inability of an autonomous system to perform predictably and reliably. However, such operations are also prone to human error due to their dynamic nature, but a different picture emerges when looking at missions that prosecute static targets, like the Chinese Embassy mission discussed earlier. These missions are typically planned days, if not weeks in advance with teams of analysts and weapons specialists evaluating multiple courses of action. Unlike the pilot in a dynamic targeting scenario, the static target pilot is not the person who picks the target (this is typically a team of people), nor is the person who authorizes the target (which is a senior official who consults with a team of lawyers). The pilot of the plane that bombs a predesignated target simply confirms that the weapons release occurs via a prescribed set of rules of engagement. For these targets, the most meaningful human control likely comes at the point of final target approval.

Indeed, the job of the pilot for these missions is primarily to get the weapon to the point in the sky where it can be released. The pilot can execute some judgment and call off a mission if various parameters are not met, such as civilians in the area. Yet this is problematic because workload is extremely high for a pilot in any weapons release scenario. In addition to making sure the plane is in the correct location, the pilot needs to consider any enemy activity such as surface-to-air missiles which is very stressful. This combination of high workload and stress also raises the likelihood of human error. So, is this kind of operation truly meaningful human control? Because of these issues and improvements in technology, this job can be and has been successfully replaced by a smart missile in many instances, suggesting that role allocation has shifted from the human to the weapon because of advancements in technology and also the relative low uncertainty of the missions.

From over 1000 miles away with meter-precision accuracy, the Tomahawk missile, a hybrid automated/autonomous technology, can deliver a similar payload to that delivered by a pilot who is susceptible to fatigue and potentially being shot down. Because of its digitized scene mapping, a form of computer vision that is highly reliable and does not use machine learning, the accuracy of the Tomahawk missile is extremely high, much better than a pilot. In the case of targeting static buildings and other targets of similar certainty (like bridges, roads, etc. that are well mapped), today's computer targeting technology far exceeds the capability of humans, especially those in a cockpit close to the action. This precision is what drives the term "surgical strike" and is why the United States prefers to use these weapons instead of humans. However, such missions are very labor and time intensive in the planning stages and the missiles cost ~$1M per copy.

Some weapons that are launched against static targets have cameras that display imagery until the point of impact, which gives the pilot a small window of time to potentially update the aim point, or even redirect a weapon off a target if some problem emerges such as the realization that the wrong target is about to be bombed due to wrong coordinates. This could be another

definition of meaningful human control. However, all the same problems exist in that these situations are extremely time-pressured, so decision biases, stress, and the fear of physical threat still can dramatically influence pilot performance. Drone missions help alleviate this problem by removing the human from direct physical threat. However, decision biases and the inability for humans to reliably assimilate all relevant information in a short period of time also affects these operators. Putting military operators into a crucible of time pressure, overwhelming volumes of information, and life and death decisions in the fog of war is very far from meaningful human control.

In short, autonomous offensive weapons already exist now that can reliably hit static, predesignated and well-mapped targets. They are also fire-and-forget weapons, which are beyond MHC, at least in real-time. They have replaced error-prone humans and absent human targeting errors, there is a high probability of success. Thus, any ban on such technology rejects not only our current use of such technologies, but also the possibility that future advances could result in autonomous offensive weapons that make fewer mistakes than humans.

Such advances are not near-term and will be rife with difficulties. The next section discusses a proposed different path that seeks to increase accountability for both engineers and military decision makers in the design and operation of these systems, as well as improve the safety of autonomous systems in general.

Focus on meaningful human certification

Whether discussing autonomous or human-controlled, weapons, MHC is ill-defined, in my opinion. The fog of war, time pressure, and the increasing amount of imperfect information available to warfighters on a battlefield all combine to make real-time decision making extremely difficult with high likelihoods of error. The most meaningful form of human control in the use of offensive autonomous weapons is deciding *a priori* which targets will be prosecuted and under what conditions.

In terms of offensive autonomous weapons, there are effectively two layers of target identification. The first is the human strategic layer, where a decision maker, typically at a high level, determines that a target (which could be a building, person, etc.) should be prosecuted in accordance with mission objectives. Such a decision must abide by principles of proportion and distinction within the Law of Armed Conflict framework. At this strategic layer, when a target is passed to lower level military personnel to prosecute, humans are certifying that the target is legitimate and the conditions of legitimacy. Obviously as in the case of the Chinese Embassy incident, mistakes are made at this strategic level, but the line of accountability is quite clear.

The second layer of target identification resides in the design of an autonomous weapon. Once a legitimate target is certified by a human decision maker, the question must be asked as to whether an autonomous weapon can correctly identify and engage a target with *better* than human odds. Such a weapon would need to adapt to changing battlefield conditions as needed to respect rules of engagement. In the case of the Tomahawk missile, if a target is static, well

mapped and within the rules of engagement, it meets this criterion. Unfortunately, only a small percentage of targets fall into this category. If dynamic and emergent targets need to be prosecuted, any autonomous weapon system used in this manner should have to prove it is *significantly better* than humans in similar circumstances. Better in this case means better at correctly identifying the target and better at responding to emergent events and respecting the rules of engagement. Unfortunately, as illustrated in Figure 1, computer visions systems that rely on machine learning techniques simply do not meet this standard at this point in time or in the near future.

This issue of autonomous system certification, i.e., that such a system can be shown to perform better than a human at a complex safety-critical task, has not been solved for either military or civilian systems, and is a topic of intense debate. Because such systems rely on probabilistic reasoning, often using black box algorithms like convolutional neural networks, we currently cannot predict with high levels of assurance how such systems will perform, especially in unforeseen conditions. Engineers do not know how to test these systems for possible errors, either errors of commission or omission. For example, engineers do not know how to predict when a computer vision algorithm will make a misidentification such as in Figure 1. Significantly more work is needed across multiple industries that wish to leverage autonomy, including not just autonomous weapons but driverless cars and surgical robots, to better understand how we can determine prior to a technology's deployment whether it truly is likely to exhibit better than human levels of performance (Cummings, 2019; Cummings & Britton, 2019).

Instead of advocating for a ban on such weapons, I believe the international community should demand that if autonomous offensive weapons are used in the future with embedded artificial intelligence, they must meet very strict certification criteria, both at the strategic layer in target selection and in the design layer with autonomous target identification. Such weapons systems should be proven through objective and rigorous testing, and should demonstrate an ability to perform better than humans would in similar circumstances, with safeguards against cybersecurity attacks. Currently in the United States, manufacturers of military weapons are indemnified against accidents on the battlefield. It is this policy that should be banned, especially for autonomous weapon systems.

Given the potential misuses and abuses of autonomous weapons technologies, the burden of proof of performance and safety should fall on the shoulders of industry as well as the military branches who buy their weapons. Both these groups should be required to certify to as-yet-determined international standards that such weapons can be used safely and responsibly, and with proof of strong safeguards against hacking attempts. Increasing accountability in this very specific way can work towards providing checks and balances needed to guard against irresponsible autonomous offensive weapon use. Exactly how to increase this accountability deserves more study as such outcomes could range from economic consequences through liability claims to a new category of war crimes.

Often those against the use of autonomous weapons will compare this technology to that of nuclear weapons, asserting that the very existence of humanity is threatened by killer robots

(Future of Life Institute, 2015). Proponents of nuclear weapon bans cite the lack of proportionality and distinction, central tenants of Just War Theory, as reasons for a ban, which are similar arguments for banning land mines. Autonomous weapons are almost the opposite in that, as demonstrated by the Tomahawk missile, they can be far more precise and proportional than most any other weapon in the existing US inventory. It may be possible, once computer vision and cybersecurity issues are addressed, for an autonomous weapon to close on a target, detect the presence of a child in the last few seconds of an attack and then divert itself to a less harmful detonation point. This is not possible for the same missile under current versions of human control.

Indeed, an argument could be made that if we know humans are prone to errors in many scenarios and are unlikely to provide actual meaningful human control, then it would be more ethical for commanders to launch a certified autonomous weapon. We have achieved this capability today for static targets (like buildings) in that it is better for a Tomahawk missile to take out a legitimate military target than it is for a human pilot. One possible derivative effect of such technologies could be reduced inclination for more diplomatic solutions, which is why a focus on the strategic certification level is just as important as the design certification level. Autonomous weapons undoubtedly can increase the pace of conflict, so from a policy perspective, slowing down the decision to use an autonomous weapon is critical as is ensuring that multiple stakeholders who have direct lines of accountability are involved.

The possibility that non-state actors could abuse such a technology is often raised as another reason for a ban. However, just as for nuclear weapons and land mines, formalized bans do not stop rogue actors. Moreover, it is not clear how a ban would stop non-state actors as anyone *today* with access to the internet and experience in a programming language like Java or C can download various free autonomous navigation, perception and face recognition programs that can be relatively easily modified to create a lethal autonomous weapon. The democratization of the internet has also, unfortunately, led to the democratization of autonomous weapons, which is what makes these systems fundamentally different from other technologies requiring special equipment and expertise.

Conclusion

Given the increasing complexity of warfare and reduced time scales, we need to recognize that we are often asking warfighters to make life and death decisions that they are ill-suited for, with high probabilities of errors. No combatant who kills anyone accidentally, military or civilian, ever truly recovers from psychological suffering, as well as the families of those killed. Autonomous weapons technology for dynamic environments will not be ready for safe or effective deployment for many years to come. However, we owe it to both those innocent people that could be killed by human error and the pilots likely to make mistakes in such settings to keep working towards a balanced role allocation in weapons systems that at least make fewer mistakes than humans. The key to this future is meaningful human certification of autonomous weapons, not insisting on an illusory concept of meaningful human control.

Acknowledgements

Deborah Johnson, Jason Borenstein, Charles Risio, and the anonymous reviewers provided feedback that substantially improved this paper.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mair, L., Ku, W. S., & Nguyen, A. (2018). Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. *arXiv: 1811.11553*.

Bibi, G. (2018). Implications of Lethal Autonomous Weapon Systems (LAWS): Options for Pakistan. *Journal of Current Affiars, 2*(2), 18-41.

Cummings, M. L. (2004). *Automation Bias in Intelligent Time Critical Decision Support Systems.* Paper presented at the AIAA 3rd Intelligent Systems Conference, Chicago.

Cummings, M. L. (2017). Artificial Intelligence and the Future of WarfareInternational Security Department and US and the Americas Programme. London: Chatham House.

Cummings, M. L. (2019). Adaptation of Licensing Examinations to the Certification of Autonomous Systems. In H. Yu, X. Li, R. Murray, S. Ramesh, & C. J. Tomlin (Eds.), *Safe, Autonomous and Intelligent Vehicles* (pp. 145-162). Basel, Switzerland: Springer International Publishing.

Cummings, M. L., & Britton, D. (2019). Regulating Safety-Critical Autonomous Systems: Past, Present, and Future Perspectives. In R. Pak, E. d. Visser, & E. Rovira (Eds.), *Everyday Robotics*.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., . . . Song, D. (2017). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint 1707.08945*.

Future of Life Institute. (2015). Autonomous Weapons: an Open Letter from AI & Robotics Researchers. Retrieved from http://futureoflife.org/open-letter-autonomous-weapons/

GAO. (1997). *Operation Provide Comfort: Review of Air Force Investigation of Black Hawk Fratricide Incident*. (GAO/OSI-9804). U.S. Washington D.C.: U.S. Government Accounting Office.

Jagacinski, R. J., & Flach, J. M. (2003). *Control Theory for Humans: Quantitative Approaches to Modeling Performance*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv:1801.00631*.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors, 52*(3), 381-410.

Rasmussen, R. E. (2007). *The Wrong Target.* Joint Forces Staff College, Norfolk, VA.

Ratches, J. A. (2011). Review of current aided/automatic target acquisition technology for military target acquisition tasks. *Optical Engineering, 50*(7). doi:10.1117/1.3601879

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition.* Paper presented at the ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria.

Wagner, M. (2014). The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems. *Vanderbilt Journal of Transnational Law, 47*, 1371-1424.