**Chapter: "War and Technology: Should Data Decide Who Lives, Who Dies?"**

by Prof. Shannon E. French
Inamori Professor in Ethics
Case Western Reserve University
General Hugh Shelton Distinguished Chair in Ethics
Command and General Staff College Foundation

When the ethical implications of the use of artificial intelligence (AI), machine learning (ML), or other forms of computing-based augmentation of decision-making are discussed in the military context, the conversation more often than not centers on lethal autonomous weapons systems, or LAWS. While this is of course an interesting area - and one I will address in this chapter - there has already been considerable attention paid to it by scholars in military ethics and international law. Since LAWS should not be the only focus of ethical analysis when it comes to the deployment of emerging military technology, in this chapter I will begin by looking at other ways in which new technology - especially technology that deals with rapid data processing - may have life-or-death consequences in modern combat and that raise other related ethical concerns. In particular, I will examine the effect of such technology on the movements of troops and equipment, including routing and navigation, on military medical triage, and on distinction (identifying possible threats, ruling out noncombatants as targets and flagging other non-targets, clearing areas as safe or hostile, surveillance in advance of troop movement, etc.). Each of these areas has its own positive opportunities technology may provide, along with its own associated ethical risks. Then I will relate these to the looming issue of LAWS.

First, I need to make a broad point about artificial intelligence. The term "artificial intelligence" is inherently misleading, as it suggests that what AI systems do is closely akin to the workings of human intelligence. This is not the case. The state of the art is nowhere near producing artificial general intelligence (AGI) that would resemble how humans think. What

actually happens in AI systems now is rapid and complex data processing, usually to seek out and identify patterns. These patterns, however, may not be what we expect. Berenice Boutin of the Asser Institute explains this point well with a simple example comparing human and machine recognition of turtles. A young child can be shown cartoon images of turtles in picture books and still go on to recognize a real turtle in a zoo or in the wild as a turtle. Current data-driven AI systems cannot make the same leap.

Human brains do use pattern recognition to reason in some contexts, but the exact mechanism of how this works is not fully understood and is the subject of debate across many disciplines, from philosophy to cognitive science. AI systems are trained to identify items by being fed examples (coded data sets) of objects or patterns until they can successfully pick out fresh examples of the same category of items. This extension from the original data to new examples depends on factors that can be coded in machine language - importantly, it is not based on anything like a form or kind, to use terms from the philosophy of language and epistemology. Willard van Orman Quine is one example of a philosopher who endeavored to explain how human inductive reasoning occurs in part due to our capacity and tendency (innate or socialized) to see "natural kinds" in the world and to shape our language to reflect these categories, whether or not they exist in reality (an issue I will certainly not attempt to resolve here). Philosophers and cognitive scientists from various schools of thought debate how classifications occur in people's minds and the exact role of language in cognition, but however this occurs, and however vague or irreducible the details may be, we somehow perceive similarities that allow us to make mental leaps like that from a cartoon turtle to the living animal.[1] We grasp "turtleness."

AI systems do not recognize "turtleness" in the same way humans do. They are trained to detect patterns in data, which is an altogether different way of chopping up the world than is

found in ordinary human cognition. It is, in a sense, profoundly alien to us. This is why MIT researchers were able to make Google AI mistake a clear picture of a tabby cat for a bowl of guacamole - or, even more relevantly in the context of military technology, a turtle for a rifle - just by manipulating pixel colors and density in an image or an object[2].

> The problem is that although neural networks can be taught to be experts at identifying images, having to spoon-feed them millions of examples during training means they don't generalize particularly well. They tend to be really good at identifying whatever you've shown them previously, and fail at anything in between. Switch a few pixels here or there, or add a little noise to what is actually an image of, say, a gray tabby cat, and Google's Tensorflow-powered open-source Inception model will think it's a bowl of guacamole. This is not a hypothetical example: it's something the MIT students, working together as an independent team dubbed LabSix, claim they have achieved.[3]

In other words, these systems are not looking for cats, or rifles, in the way that humans would. It is a mistake to anthropomorphize artificial systems as if they were just super-intelligent people, doing what we do, only faster and better. It may be more challenging to guard against the kinds of errors a machine would make, as opposed to the all-too-familiar failings to which people fall prey.

How AI classifications work is a "black box" problem. It can be difficult to gauge how trustworthy a classification system is by looking at it from the outside, as it were, because results that may initially seem to be accurate may be based on qualities that are fundamentally unreliable. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin demonstrated this point quite effectively with an experiment that showed how a bad classifier could pick out pictures of wolves from pictures of huskies most of the time, simply by classifying pictures with snow in the background as wolf pictures and those with grass behind the animal as dog pictures:

> Often artifacts of data collection can induce undesirable correlations that the classifiers pick up during training. These issues can be very difficult to identify just by looking at the raw data and predictions. In an effort to reproduce such a setting, we take the task of distinguishing between photos of Wolves and Eskimo Dogs (huskies). We train a logistic regression classifier on a training set of 20 images, hand selected such that all pictures of wolves had snow in the background, while pictures of huskies did not. As the features for the images, we use the first max-pooling layer of Google's pre-trained Inception neural network [25]. On a collection of additional 60 images, the classifier predicts "Wolf" if there is snow (or light background at the bottom), and "Husky" otherwise, regardless of animal color, position, pose, etc. We trained this bad classifier intentionally, to evaluate whether subjects are able to detect it.[4]

Even though the subjects of the experiment were "graduate students who have taken at least one graduate machine learning course," far too many of them nonetheless trusted the bad classifier, until the patterns on which it had been trained were fully explained to them. Their faith was misplaced, and more robust than it should have been. As the experimenters sagely observe, "Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic."[5]

It takes very careful coding and training of classification algorithms to ensure that they are reasonably reliable, and even the best systems remain vulnerable to "adversarial examples," or images intended to trick the classifier. Currently, the most effective systems are those that are able to analyze fairly static, consistent images, with as few variables as possible to interfere with correct classifications. For example, there appear to be genuine benefits from AI systems reviewing mammogram images to look for breast cancer.[6] Unfortunately for those who support military applications for AI classification systems, armed conflict involves ever-changing

conditions that do not lend themselves well to automated analysis, combined with the near certainty of adversarial efforts to intentionally push such systems into false classifications.

Why does this matter? After all, humans can also be fooled and manipulated and make identification errors or pattern recognition mistakes. It matters primarily because people have been exposed to quite a lot of "sales pitches" suggesting that AI can do more than it can (or will be able to do for decades, if ever). It is vital not to fall for hype about the capabilities of AI/ML systems for a number of reasons. For one thing, there is the strong danger of automation bias. Automation bias is the tendency to accept the authority of an automated or computerized system as superior to one's own or another human's authority - to assume that the machine knows best. As Lisa Lindsay and I have noted elsewhere, this can be a gravely pernicious aspect of the use of automated systems by the military:

> Artificial intelligence, as a technology that is little understood by the general public and often sold as 'better' than human ability, puts us at even higher risk for automation bias. Despite documented failures, people focus on reports that seem to suggest that artificial minds are superior to organic ones, including news of computers beating humans at strategy games like chess and Go. While television advertisements find humor in people following incorrect Google maps directions straight into a lake, the reality of overreliance on automation is much less amusing. In a military context, automation bias can have life or death consequences. Just as infantry check the proper functioning of their weapons, those using A.I. systems in their military roles are obligated to make sure their tools – both automated and not – are working correctly, too.[7]

Elke Schwarz warns that, "[s]et against a background where the instrument is characterized as inherently wise, the technology gives an air of dispassionate professionalism and a sense of moral certainty to the messy business of war."[8] While those most critical of the military tend to focus on instances of human error or, worse, intentional evil, examples also abound of human

troops showing restraint, compassion, and kindness, even at great risk to themselves. I am reminded of this case, from the US's engagement in Somalia in 1992:

> Colonel Michael Campbell, another US Marine, was commanding a tank unit in Somalia when he was ordered to destroy three tanks that the enemy had deployed on the outskirts of an impoverished civilian community. When Colonel Campbell's unit, which included both armour and infantry, came into range, the turrets of the enemy tanks turned towards the approaching American troops. The colonel's subordinates in the US tanks urgently requested permission to fire first, to defend themselves and the infantrymen all around them. However, something made the colonel hesitate. Perhaps it was just instinct, or perhaps there was something slightly wrong about the way the enemy turrets turned to bear upon the Americans that set off alarm bells in the colonel's subconscious. For whatever reason, with both his superiors on the radio and his subordinates all around him shouting for him to order the attack, the colonel refused to fire at the tanks. Just then, the hatches on top of the enemy tanks popped open, and Somali children began to crawl out and run back to their homes. The tanks had been abandoned in the middle of the night, and the children had been playing in them.[9]

Imagine how much additional pressure the colonel might have felt to order fire if there had been an AI-enabled targeting system firmly identifying the tanks as urgent military targets. Whether we should trust machine systems more than human systems to make final decisions on use of lethal force is a topic I will return to later when I discuss LAWS, but for human-in-the-loop military decision-making, the risk of individuals allowing automation bias to cloud their judgment must be given proper weight. There is disquieting research that suggests the possible harms that could follow from the effects of automation bias. MIT's M.L. Cummings summaries some examples of this in the context of automated route-planning systems:

> [C]omputer generated solutions are not always truly optimal and in some cases, not even correct. Known as "brittleness," automation decision support models in complex systems cannot account for all potential conditions or relevant factors which could result in erroneous or misleading suggestions.[1011] In addition, as problem spaces grow in complexity, it becomes more difficult for the human to not only understand whether or not a computer-generated solution is correct, but how any one variable, a combination of variables, or missing information influence the computer's solution. This inability of the human to understand complex algorithms only exacerbates the tendency towards automation bias. For

> example, in a study examining commercial pilot interaction with automation in an enroute flight planning tool, pilots, when given a computer-generated plan, exhibited significant automation over-reliance causing them to accept flight plans that were significantly sub-optimal. When presented with an automated solution, 40% of pilots reasoned less or none at all when confronted with uncertainty in the problem space and deferred to erroneous automation recommendations, even though they were provided with tools with which to explore the automation space. The authors of this study suggest that even if automated critiquing alerts are provided to warn against possible constraint violations and/or provide suggestions to avoid constraint violations, human decision makers can be susceptible to automation bias 16. In a similar experiment looking at levels of automated assistance in a military in-flight replanning task, pilots with [automated] assistance exhibited overreliance in the form of complacency. In this study, an automated decision aid planned a route taking into consideration time to targets, possible threats, and fuel state and subsequently presented pilots with its "optimal" solution, which could always be significantly improved through human intervention. Despite having the ability to change and improve the computer's solutions, subjects tended to accept the computer's solution without question.[12][13]

From his extensive research, Cummings concluded that, "in time critical environments with many external and changing constraints such as air traffic control and military command and control operations, higher levels of automation are not advisable because of the risks and the complexity of both the system and the inability of the automated decision aid to be perfectly reliable," further noting that, "there can be measurable costs to human performance when automation is used, such as loss of situational awareness, complacency, skill degradation, and automation bias."[14]

There are other forms of bias to contend with, as well. Systems that are programmed by humans, even if those systems proceed to "learn" on their own, are vulnerable to the same biases and fallibility of humans. There is an essential truth captured in the phrase, "garbage in, garbage out." A grimly humorous example of this can be seen in the case of an AI "chatbot" named Tay that Microsoft designed to interact with people on Twitter. The intention was for the chatbot to acquire the ability to mimic human interactions more naturally. The result was that Tay quickly began to send out wildly offensive tweets, including all forms of hate, including (but not limited

to) racism, misogyny, and anti-Semitism.[15] In a similar case, 'after Facebook eliminated human editors who had curated "trending" news stories... the algorithm [that replaced the human editors] immediately promoted fake and vulgar stories on news feeds.'[16]

Automated systems are certainly not amoral or divorced from the ethical or character flaws of humans. In *Weapons of Math Destruction*, Cathy O'Neill points out, "these models [of AI and ML] are constructed not just from data, but from the choices we make about which data to pay attention to - and what to leave out. Those choices are not just about logistics, profits, and efficiency. They are fundamentally moral."[17] Even as such systems are improved, they need to be seen as tools that process information only in very specific ways, not as superhuman or godlike intelligences that are necessarily more accurate or objective than human agents. In other words, we need to understand going into any discussion about the ethical use of AI/ML systems that these systems do not think like us and that they can only work with the information (data) that we give them. With this background perspective in mind, let us review some of the ethical issues and the pros and cons that arise when deploying data-driven systems for non-lethal military use.

First, consider the case of military medical triage. As Sara Gaines has explained[18], the US military, like most, divides military medical care into three categories: (1) Care Under Fire (CUF); (2) Tactical Field Care (TFC); and (3) Tactical Evacuation Care (TACEVAC). All three of these collectively are known as Tactical Combat Casualty Care (TCCC)[19]. The idea behind AI-augmented military medical triage would be to have an algorithmic, data-based system to help medics on the ground make time-sensitive triage decisions. Gaines explains the complex context in which this technology could be deployed:

> Moral injury can be defined as a "disruption in an individual's confidence and expectations about one's own or others' motivation or capacity to behave in a just and ethical manner."[20] During TFC, the medics making decisions have their dual loyalties to contend with, meaning that there are times when one decision may go against the expectations of the other role. In the scenarios mentioned before, a medic who follows orders and does not treat a casualty is acting as a soldier, but there is then the psychological cost of not acting within the role of care provider to the casualty. For combat medics who are expected to follow international laws regarding treatment of casualties, they need to be able to have faith in themselves to make the correct decision. For medics suffering from a moral injury and lacking the ability to trust their own decisions, making the tough in-arena calls which must be made would be extraordinarily difficult. If a medic is not able to make the difficult decisions associated with the role and act quickly, then it is the casualties who must incur greater suffering until treatment begins. Not only are there in-arena consequences to a medic suffering from a moral injury and unable to trust that the decisions being made by them and around them are the just and ethical ones, but suffering a moral injury makes them more susceptible to developing Post-Traumatic Stress.[2122]

In theory, an AI system might spare human medics some degree of moral injury by taking the most gut-wrenching decisions out of their hands. Would that really happen, though, or would medics still second-guess whether they should have, for example, overridden the system's suggestions? I share Gaines' skepticism both that we are anywhere close to having the technological capability to field an AI triage system and that such a system would reduce moral injury for medics:

> [E]ven if the medic believes the AI to be an accurate tool, the possibility of a moral injury still remains. Evidence has developed in research on post-traumatic stress that there is often guilt associated with not being subject to the same harm as those who went through the traumatic event as well.[23] Known as survivor guilt, one important component within the context of a combat medic is that this guilt can develop from a sense of feeling as though a different decision could have or should have been made, even when no other option existed.[24] So even if an AI is capable of making the decision, unless the medic fully agrees with that decision, they are still at risk of their loyalty as a soldier conflicting with their loyalty as a medic to care for the casualties around them.[25]

This raises another ethical concern: what does healthy dissent look like when AI systems are given a role in decision-making, either as an advisor or an authority? Again, extensive research on automation bias proves that, consciously or unconsciously, humans often tend to cede authority to automated systems and assume them to be more objective or fair.

Unfortunately, as Safiya Noble unequivocally demonstrates in *Algorithms of Oppression*, it is just as likely such systems will be as much or more biased than their human counterparts, and will amplify that bias by imbuing it with the imagined infallibility of cold machine logic. As Noble states, "Algorithms are, and will continue to be, contextually relevant and loaded with power."[26] In civilian medicine, we have already seen how systems trained on, for example, non-diverse data sets that do not contain data from different races, can have a negative impact on patient outcomes if doctors rely on them alone or invest them with too much authority. Interestingly, a recent study revealed that senior medical professionals were much more likely to trust their own judgment than to take the AI system's advice, whereas less experienced doctors tended to second-guess themselves and defer to the machine.[27] This has implications for a military setting, where differences in experience and rank may either exacerbate or mitigate automation bias. Do we want human medics to "go with their gut," knowing that their "gut" may be honed from years of experience or subject to irrational impulse? Or do we want them to let the data decide who can still be saved?

One way to frame the issues is to ask where is the threshold to know when it is better to let the data decide who lives, who dies? Is there a perfect percentage point where we should feel morally comfortable saying "go with the machines," because, for instance, the machines make X percentage fewer errors than humans do in similar circumstances? This way of approaching the

problem, by comparing error rates, is tempting. Yet it may be fundamentally misguided. It matters ethically *what kinds of errors are made*, not just how many errors there are. For example, an AI triage system that makes fewer overall errors than human medics do but that consistently underestimates the survival potential of a particular gender or race would not be ethical to deploy. To give an everyday example, a few seasons ago, my child's summer camp offered a supposedly AI-enabled system using facial recognition to pick out camp photos for parents that contained their children. Setting aside other ethical concerns with this, I noticed something interesting about the errors that the system made. In one large group of photos, it picked out six that it flagged as containing my child. Of these, four indeed did contain my child, but one did not contain any children at all (only trees and other objects). Another was of a child that was not only not mine but in no way resembled mine. In other words, these were not only errors, they were errors that no reasonable person could have made. Meanwhile, I found a total of 61 photos that contained my child, and was able to accurately discern from subtle cues in some of them that she was unwell, which was soon confirmed by the camp. The point of this unscientific case study is that we must approach claims of lower-than-human error rates for automated systems with some skepticism, and ask the right questions.

Let us turn now to another possible military application for automated systems, to see if the same, or different, ethical issues arise. Initially, the movement of troops and equipment (logistics) and routing and navigation seem like areas that would be less fraught with ethical peril than military medical decision-making. The benefits also seem fairly clear, if we consider for instance the appeal of just-in-time deliveries of correct needed parts to units, made possible by an advanced automated system capable of machine learning that tracks and learns where supplies will run low when. It is also straightforward to grasp the advantages of routing systems that give

a commander real-time options to navigate through unfamiliar terrain, updated with more data than human scouts could ever hold in their minds. There is positive potential here to build and increase capacity and efficiency, but once again, the devil is in the details.

As any cyber security expert will tell you, even the most seemingly autonomous system remains susceptible to human error, as well as to malicious human attack. A just-in-time supply chain depends on correct, updated information. Once again, it is garbage in, garbage out. Any automated system can be hacked and undermined or simply fall prey to mistakes or unanticipated elements of chaos introduced by non-human factors like the environment or human ones.[28] This is especially true of systems that depend on complex digital platforms to collect and communicate information.

> Even as they grant unprecedented powers, [digital technologies] also make users less secure. Their communicative capabilities enable collaboration and networking, but in so doing they open doors to intrusion. Their concentration of data and manipulative power vastly improves the efficiency and scale of operations, but this concentration in turn exponentially increases the amount that can be stolen or subverted by a successful attack. The complexity of their hardware and software creates great capability, but this complexity spawns vulnerabilities and lowers the visibility of intrusions. Cyber systems' responsiveness to instruction makes them invaluably flexible; but it also permits small changes in a component's design or direction to degrade or subvert system behavior. These systems' empowerment of users to retrieve and manipulate data democratizes capabilities, but this great benefit removes safeguards present in systems that require hierarchies of human approvals. In sum, cyber systems nourish us, but at the same time they weaken and poison us.[29]

Every new piece of technology introduces new potential points of failure. As Jacquelyn Schneider explains, this can be understood as the "Capability-Vulnerability Paradox," and it is not a unique problem for digital or computerized systems:

> In examining analogies within infrastructure development and conflict, a historical pattern of capabilities and vulnerabilities that illustrate the logic of the capability/vulnerability paradox emerges. Take, for example, the combustion engine. Internal combustion engines opened up remarkable opportunities for

> weapons development – from tanks to aircraft to ships, combustion engines made nations more effective on the battlefield. But it also made them more dependent on oil and therefore vulnerable to disruptions in the oil supply chain.[30]

Some of these concerns are practical, not ethical, and as such are not necessarily grounds to refuse to develop and continuously improve such systems. However, ethics will quickly come into the picture if militaries allow themselves to become over reliant on such systems without establishing appropriate back-ups and potential overrides. In the tragic Boeing 737 Max airplane crashes, when the automated systems went wrong, the human pilots found themselves unable to seize back manual control and save the lives of their passengers (or themselves).[31] The automation itself blocked recovery of the aircraft. These are design choices that have ethical consequences.

> The U.S. military has a troubling history of implementing systems without ample time for them to be carefully studied and tested from a safety perspective, let alone from a legal or ethical standpoint. The Bradley fighting vehicle and the osprey are just two well-known examples of flawed systems rushed into use. There are also the tragedies of service personnel who were sent into irradiated areas before the effects of nuclear weapons were understood or the combat troops who were given unreliable, jam-prone M-16s in Vietnam: 'from Gettysburg to Hamburger Hill to the streets of Baghdad, the American penchant for arming troops with lousy rifles has been responsible for a staggering number of unnecessary deaths.'[3233]

Safety testing is only part of the picture. Great care must be taken before militaries buy into systems that may not only not have humans in the loop but may actively lock humans out of the loop even, or especially, when things go wrong with the system itself. That is unacceptable. There is a reason why the NASA astronauts of the Mercury program strongly objected to complete automated control of windowless capsules, which famous test pilot Chuck Yeager said reduced them to mere "spam in a can."[34]

Progress has been made in this area, and there are hopeful signs, such as military research funding organizations like DARPA increasing their requirements for on-going ELSI/LME

(Ethical, Legal, and Social Issues/Legal, Moral, and Ethical) reviews of developing projects that are not mere check-the-box compliance exercises. Nevertheless, legitimate concerns persist in light of the U.S. military's tendency to characterize every potential technological advancement as an urgent upgrade that must be deployed as quickly as possible to gain an advantage. This "arms race" attitude is not only reckless, it also fails to take into account not only possible harms from improperly vetted systems but the uncomfortable truth of asymmetric conflicts. It is simply not the case that the more technologically advanced side in an asymmetric armed conflict always (or even usually) prevails. Nor is it consistently true even in peer or near-peer conflicts that the first side to deploy a particular technological advancement always gains the advantage. Sometimes, the second mouse gets the cheese.

The just war tradition (JWT) immediately becomes relevant when we turn to looking at the use of AI/ML systems to attempt to distinguish between combatants and noncombatants. Here it is especially important to remember the point I reviewed at the start of this chapter concerning how these systems "recognize" things and spot patterns. It is a value-laden question to ask, if you want to train a computerized system to determine if a person is or is not a legitimate target in war, what exactly should you tell it to look for - bearing in mind that you cannot rely on human-like cognition and understanding of kinds? Should you try to train it to look for a weapon? What do weapons look like, to a pattern-analyzing machine? What would reliably distinguish a rifle from other objects? What about cruder weapons? What about inherently inconsistently constructed weapons like improvised explosive devices (IEDs)? Humans find these kinds of identifications challenging, too, especially when under extreme stress. Would a system of pattern recognition do better? As before, we have to ask what kind of error rates we need, and are certain types of errors more or less ethically tolerable than others?

For example, is mistaking a child's toy for a gun worse than mistaking a carton of cigarettes for an IED?

Suppose we decide that looking for weapons seems too problematic. The alternatives might be even worse, since they would most likely involve trying to determine combatant/noncombatant status by an individual's hostile intent (or lack thereof). Exactly how would you train a system to pick out hostile intent? It is notoriously difficult to predict and analyze human behavior and responses, and as Ruha Benjamin points out in *Race After Technology*, these issues are even harder when dealing cross-culturally, with diverse races, genders, ages, and communities, all in high-stress circumstances likely themselves to skew "normal" behavior.[35] In testifying before Congress about the extensive failures of the Transportation Security Agency (TSA)'s Screening of Passengers by Observation Techniques (SPOT) program, psychologist Phil Rubin noted how the underlying "science" behind using things like machine-detected "micro-expressions" to determine potential hostility was nothing but false "snake oil," with no reliable, verifiable basis: "In our desire to protect our citizens from those who intend to harm us, we must make sure that our own behavior is not unnecessarily shaped by things like fear, urgency, institutional incentives or pressures, financial considerations, career and personal goals, the selling of snake oil, etc., that lead to the adoption of approaches that have not been sufficiently and appropriately scientifically vetted."

What we would want from an ethics perspective would be a system that could assist human troops with discrimination in a way that would nudge more towards erring on the side of assuming someone is a noncombatant - that would focus on helping to prevent wrongful targeting and deaths. There are many reasons to hope for the development of such a system, not

the least of which is concern for the well-being of the troops themselves, who, as I have argued extensively elsewhere[36], can suffer moral injury when targeting mistakes are made or collateral damage assessments are incorrect. Cases of troops intentionally committing war crimes against civilians are thankfully rare, but tragic mistakes, including also the category of "friendly fire" incidents, are more common. Assisting troops with avoiding killing those who do not need to die to achieve the mission would be a goal worth achieving. Sadly, though, systems that predict or suggest possible hostility based on bad data or weak (hard or social) science could do more harm than good, leading to more suffering for all those concerned, including the troops.

Last but not least, we cannot ignore the reality that as soon as new technology is fielded, the people exposed to it will quickly adapt or evolve their tactics in response. This is ethically relevant because it highlights a sub-species of automation bias. If troops are assured that they have been given systems that can identify hostile forces, but those systems are learned and fooled by their enemies after the first or second use, they may be more a burden than a boon. People are creative and resourceful. A new system may yield a temporary advantage, but as soon as it is known, its vulnerabilities will be exploited. Where there are humans, there is chaos, and unpredictability is an advantage that does not require a computer.

Up until this point, we have been dealing with systems that have no executive function; where targeting and distinction decisions can be overridden by humans. Now we enter a different world, where such decisions are entirely synthetic in agency - the realm of lethal autonomous weapons systems (LAWS). The core ethical argument in favor of the use of LAWS depends on the claim that such systems present a possible solution to the problem of uniquely human qualities (including character deficiencies, natural reactions to extreme stress, or other psychological, biological, or moral factors) leading to either tragedies or intentional war crimes.

One of the strongest proponents of this view is Ronald Arkin, who provides ethics consultation for engineering research and development projects aimed at creating what others call "killer robots" - but which he believes could be better stewards of legal and ethical conduct in war:

> Defending the use of robots and other automated systems in the military, Ron Arkin has essentially argued that humans are too emotionally vulnerable to be trusted to do the right thing in combat conditions. Citing surveys in which military personnel admit to unethical views about the importance (or lack thereof) of obeying the laws of war, Arkin asserts that humans are too often overcome by intense feelings such as rage and fear (or terror) that effectively hijack their brains and can lead even to the perpetration of war crimes. In his book Governing Lethal Behavior in Autonomous Robots, Arkin avers that, '...it seems unrealistic to expect normal human beings by their very nature to adhere to the Laws of Warfare when confronted with the horror of the battlefield, even when trained.'[37] He believes robots can do better.[38]

Arkin's claim that robots would do better is difficult to test or refute so long as he is comparing imperfect humans to theoretical future robots. Many others in military ethics, myself included, are unconvinced by the argument that human troops are doomed always to fall into the trap of crossing ethical lines under the strain of combat. While violations of the letter and spirit of the Law of Armed Conflict (LOAC) do occur, with more regularity than anyone would like, at least within regular, well-disciplined units, they are certainly the exception, not the rule. Careful analysis of the causes and contributing factors of such incidents, as is found in Jessica Wolfendale and Matthew Talbert's *War Crimes: Causes, Excuses, and Blame*[39], for example, and insights on the causes and effects of dehumanization in war as explored by David Livingstone Smith in *Less Than Human* and *On Inhumanity*[40], can provide a roadmap for enhanced training and education of troops and their leadership to reduce or perhaps even prevent these occurrences. Meanwhile, there are strong ethical arguments against shifting the responsibilities of killing in war onto machines.

As Elke Schwarz points out, the use of fully autonomous lethal machines in war would damage or severely restrict the proper attribution of ethical responsibility and accountability by creating a "moral vacuum" where crucial decision-making cannot be traced back to any moral agent. This is especially true in cases (which are common in armed conflicts) where guidance from existing laws and norms is inconclusive:

> What I wish to highlight here is the moral vacuum that technologies of ethical decision-making create in their quest to 'secure' moral risk. A moral vacuum opens when certain parameters of harm are no one's responsibility; when the decision that harm is permissible has been determined through technological means. This moment is, paradoxically, also the very moment of moral responsibility. In other words, the moral vacuum exists exactly in the moment when neither law nor existing moral guides have adequate reach. It is in this moment where responsibility resides. For example, the moral vacuum opens precisely when a specific signature strike is executed. Here the decision to kill someone has been determined by algorithms that feed into a disposition matrix and determine whether an individual's pattern of life analysis betrays terroristic dispositions. The accuracy of the technologically determined kill decision cannot easily (if at all) be verified. Where, then, resides responsibility?[41]

In *Technology and the Virtues,* Shannon Vallor shares a similar ethical unease with the idea of humans off-loading life-or-death decision-making to machine systems. She argues further that such a shift of responsibility would have a deleterious effect on human community and the human exercise of essential virtues:

> The oft-promised ability of drone warfare to minimize civilian casualties from airstrikes has yet to be empirically demonstrated by any neutral observer, but if the asymmetry that modern warfare fosters plays *any* role in feeding the warped psychology and recruitment successes of groups like ISIS and Al Qaeda, then the claim that robotic warfare will make innocent civilians overall safer from the horrors of war is plainly dubious. Thus the development of lethal military robots that promise to allow a minority of privileged human beings to detach even further from the physical, psychological, and moral horrors of war endured by the rest of humanity is deeply inconsistent with the technomoral virtue of courage, not to mention justice, empathy, and moral perspective.[42]

The principle of distinction is the central tenet of jus in bello. Without that core determination of who is or is not a legitimate target, armed conflicts rapidly descend into the utter horror of indiscriminate slaughter. In "59 Percent Likely Hostile," Daniel Eichler and Ronald Thompson give the example of a system of discrimination similar to those being explored now with funding from organizations such as DARPA, "The application warns [the soldier] that a group of three adult males are one kilometer ahead and closing. It assigns them a 59 percent chance of being hostile based on their location, activity, and appearance. Clicking on this message, three options flicker onto his screen – investigate further, pass track to remotely piloted aircraft, or coordinate kinetic strike."[43] Eichler and Thompson acknowledge the concern that the psychological effect of automation bias and a lack of statistical understanding may drive the soldier to treat the unknown men as definitely hostile and engage them with deadly force without further reflection or evidence. They go on to argue that the remedy for this is to train future troops in a better understanding of how both statistics and algorithms work, so that they will be less likely to leap to conclusions or misinterpret results. This seems to me to miss a critical point. There is an ethically relevant difference between a system that provides additional information (e.g. "their location, activity, and appearance") from which soldiers make their own determination of the possible hostility of unknown persons and one that actually labels those persons – even with the "hedge" of an assigned percentage – as hostile or not a threat.

Ultimately, the question here is can we reduce the principle of distinction to a numbers game? If (and that is very a big "if") we can collect the right data, can machines be designed to tell soldiers whom to shoot, more reliably than their own senses? Noel Sharkey, an expert on AI and robotics and co-founder of the International Committee for Robot Arms control, is intensely skeptical of any plan to fully automate lethal targeting:

> A computer can compute any given procedure that can be written down in a programming language. We could, for example, give the robot computer an instruction such as, "If civilian, do not shoot". This would be fine if, and only if, there was some way of giving the computer a clear definition of what a civilian is. We certainly cannot get one from the Laws of War that could provide a machine with the necessary information. The 1944 Geneva Convention requires the use of common sense, while the 1977 Protocol 1 essentially defines a civilian in the negative sense as someone who is not a combatant…. And even if there was a clear computational definition of a civilian, we would still need all of the relevant information to be made available from the sensing apparatus. All that is available to robots are sensors such as cameras, infrared sensors, sonars, lasers, temperature sensors and ladars, etc. These may be able to tell us that something is a human, but they could not tell us much else. In the labs there are systems that can tell someone's facial expression or that can recognise faces, but they do not work on real-time moving people. [...] There is also the Principle of Proportionality and again there is no sensing or computational capability that would allow a robot such a determination, and nor is there any known metric to objectively measure needless, superfluous or disproportionate suffering.They require human judgement.[44]

Warfare always drives innovation, and it is only to be expected that people will look for ways to use technology to try to better survive future conflicts. From an ethical perspective, however, the type of survival that matters is more than physical. However well intended, the wrong applications of emerging tools could increase rates of moral injury amongst troops while also causing additional tangible harm to the most vulnerable populations. The incorporation of new technology into military operations must therefore be handled with great care and deliberation, not in a mad rush to be the first out of the gate. War is not a game of chess or go, nor is it readily reducible to zeros and ones. As World War II combat veteran J. Glenn Gray poignantly reminds us in *The Warriors: Reflections on Men in Battle*, "For all its inhumanity, war is a profoundly human institution." There may be ways to innovate intelligent systems that truly augment troops, but when it comes to deciding who lives and who dies, we have to keep the human in the loop.[45]

[1] Quine, Willard van Orman. "Natural Kinds." *Ontological Relativity and Other Essays*, edited by W. V. Quine, Columbia University Press, 2012, pp. 114–138.

[2] Katyanna Quach, "How we fooled Google's AI into thinking a 3D-printed turtle was a gun: MIT bods talk to El Reg," *The Register*, November 6, 2017.

[3] Katyanna Quach, "How we fooled Google's AI into thinking a 3D-printed turtle was a gun: MIT bods talk to El Reg," *The Register*, November 6, 2017.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin,"'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD '16: *Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, pages 1135-1144.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin,"'Why Should I Trust You?' Explaining the Predictions of Any Classifier," KDD '16: *Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, pages 1135-1144.

[6] McKinney, S.M., Sieniek, M., Godbole, V. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94 (2020).

[7] Shannon E. French and Lisa N. Lindsay, "Artificial Intelligence in Military Decision-Making: Avoiding Ethical and Strategic Perils with an Option-Generator Model," Bernard Koch and Richard Schoonhoven, editors, *The Ethical Implications of Emerging Technologies in Warfare,* The Netherlands and Boston: Brill/Martinus Nijhoff Publishers, forthcoming.

[8] Elke Schwarz, 'Technology and moral vacuums in just war theorising' (2018) *Journal of International Political Theory* 1.

[9] Shannon E. French, "An American Military Ethicist's Perspective: Such Waste in Brief Mortality," *The Price of Peace: Just War in the 21st Century*, Charles Reed and David Ryall, editors, Cambridge: Cambridge University Press (2007).

[10] Smith, P., McCoy, E., and C. Layton, Brittleness in the design of cooperative problem-solving systems: The effects on user performance, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27, 1997, 360-371.

[11] Guerlain, S., Smith, P., Obradovich, J., Rudmann, S., Strohm, P., Smith, J., and Svirbely, J., Dealing with brittleness in the design of expert systems for immunohematology, Immunohematology, 12, 1996, 101-107.

[12] Johnson, K., Ren, L., Kuchar, J., and Oman, C., Interaction of Automation and Time Pressure in a Route Replanning Task, International Conference on Human-Computer Interaction in Aeronautics, Cambridge, MA, 2002. 132-137.

[13] M.L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," Decision Making in Aviation (2017).

[14] M.L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," Decision Making in Aviation (2017).

[15] James Vincent, 'Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day,' *The Verge* (24 March 2016).

[16] Sam Levin, 'A beauty contest was judged by AI and the robots didn't like dark skin,' *The Guardian*, (8 September 2016).

[17] Cathy O'Neill, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown Publishing Group (2017), p.218.

[18] Sara Gaines, "Who Should Choose? Impacts of Artificial Intelligence Use in Tactical Field Care," MA candidate graduate term paper for PHIL417, "War and Morality," Case Western Reserve University, Fall 2018.

[19] Stephen D. Giebner, "The Transition to the Committee on Tactical Combat Casualty Care." *Wilderness & Environmental Medicine* 28, no. 2 (2017).

[20] Kent Drescher, David W. Foy, Caroline Kelly, Anna Leshner, Kerrie Schutz, and Brett Litz. "An Exploration of the Viability and Usefulness of the Construct of Moral Injury in War Veterans." Traumatology 17, no. 1 (2011): 9.

[21] Brett T. Litz, Nathan Stein, Eileen Delaney, Leslie Lebowitz, William P. Nash, Caroline Silva, and Shira Maguen. "Moral Injury and Moral Repair in War Veterans: A Preliminary Model and Intervention Strategy." *Clinical Psychology Review* 29, no. 8 (2009): 702.

[22] Sara Gaines, "Who Should Choose? Impacts of Artificial Intelligence Use in Tactical Field Care," MA candidate graduate term paper for PHIL417, "War and Morality," Case Western Reserve University, Fall 2018.

[23] Sadie P. Hutson, Joanne M. Hall, and Frankie L. Pack. "Survivor Guilt." Advances in Nursing Science 38, no. 1 (2015): 20-21.

[24] Sadie P. Hutson, Joanne M. Hall, and Frankie L. Pack. "Survivor Guilt." Advances in Nursing Science 38, no. 1 (2015): 27.

[25]Sara Gaines, "Who Should Choose? Impacts of Artificial Intelligence Use in Tactical Field Care," MA candidate graduate term paper for PHIL417, "War and Morality," Case Western Reserve University, Fall 2018.

[26] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press (2018), p.171.

[27] Gaube, S., Suresh, H., Raue, M. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* 4, 31 (2021).

[28] I am reminded here of the character "Crapgame" from the 1970 war movie "Kelly's Heroes," who is an apt fictional - but not unrealistic - example of a human in the mix capable of adding considerable chaos to any system trying to track military supplies.

[29] Richard Danzig, "Surviving on a Diet of Poisoned Fruit: Reducing the National Security Risks of America's Cyber Dependencies," (Center for New American Security, July 2014), accessed 28 August 2021.

[30] Jacquelyn Schneider, 'Digitally-Enabled Warfare: The Capability-Vulnerability Paradox' (Center for a New American Security 2016) accessed 28 August 2021.

[31] Dominic Gates, "Pilots Struggle Against Boeing 737 Max Control System on Doomed Lion Air Flight," Seattle Times, November 27, 2018.

[32] Robert H. Scales, 'Gun Trouble,' *The Atlantic* (January/February 2015).

[33] Shannon E. French and Lisa N. Lindsay, "Artificial Intelligence in Military Decision-Making: Avoiding Ethical and Strategic Perils with an Option-Generator Model," Bernard Koch and Richard Schoonhoven, editors, *The Ethical Implications of Emerging Technologies in Warfare,* The Netherlands and Boston: Brill/Martinus Nijhoff Publishers, forthcoming.

[34] Jannelle Warren-Findley, "The Collier as Commemoration: The Project Mercury Astronauts and the Collier Trophy," https://history.nasa.gov/SP-4219/Chapter7.html.

[35] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, Cambridge: Polity Press (2019).

[36] See Shannon E. French, *The Code of the Warrior: Exploring Warrior Values, Past and Present*, Lantham, MD: Rowman & Littlefield Publishers, second edition (2017).

[37]Ron Arkin, *Governing Lethal Behavior in Autonomous Robots,* Taylor & Francis Group (2009) 36.

[38] Shannon E. French and Lisa N. Lindsay, "Artificial Intelligence in Military Decision-Making: Avoiding Ethical and Strategic Perils with an Option-Generator Model," Bernard Koch and Richard Schoonhoven, editors, *The Ethical Implications of Emerging Technologies in Warfare,* The Netherlands and Boston: Brill/Martinus Nijhoff Publishers, forthcoming.

[39] Matthew Talbert and Jessica Wolfendale, *War Crimes: Causes, Excuses, and Blame*, Oxford: Oxford University Press (2018).

[40] David Livingstone Smith, *Less Than Human: Why We Demean, Enslave, and Exterminate Others*, New York: St. Martin's Press (2012) and *On Inhumanity: Dehumanization and How to Resist It*, Oxford: Oxford University Press (2020).

[41] Elke Schwarz, "Technology and moral vacuums in just war theorising,"*Journal of International Political Theory* 2018, Vol. 14(3) 280–298.

[42] *Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Oxford: Oxford University Press (2016) 216.*

[43] Daniel Eichler and Ronald Thompson, "59 Percent Likely Hostile," War on the Rocks, *Texas National Security Review*, January 2020.

[44] Noel Sharkey, "Grounds for Discrimination: Autonomous Robot Weapons," RUSI Defense Systems, October 2008.

[45] J. Glenn Gray, *The Warriors: Reflections on Men in Battle*, New York: Harper and Row, 1970, pps. 152-153.