Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture PART I: Motivation and Philosophy

Ronald C. Arkin
Mobile Robot Laboratory
Georgia Institute of Technology
Atlanta, GA. 30308
1-404-894-8209
arkin@cc.gatech.edu

ABSTRACT
This paper provides the motivation and philosophy underlying the design of an ethical control and reasoning system potentially suitable for constraining lethal actions in an autonomous robotic system, so that its behavior will fall within the bounds prescribed by the Laws of War and Rules of Engagement. This research, funded by the U.S. Army Research Office, is intended to ensure that robots do not behave illegally or unethically in the battlefield. Reasons are provided for the necessity of developing such a system at this time, as well as arguments for and against its creation.

Categories and Subject Descriptors
I.2.9 [**Artificial Intelligence**]: Robotics – *autonomous vehicles*

General Terms
Legal Aspects.

Keywords
Human-Robot interaction, Robot ethics, Battlefield robots, Unmanned systems, Autonomous robots.

This paper is the first in a series of papers that describe the design and implementation of an "artificial conscience" for autonomous robotic systems capable of lethal force. This project, funded by the U.S. Army Research Office (ARO) is being conducted over a three-year period. Representational choices and specific architectural components for this system appear in subsequent articles [1,2], and are also described in detail in [3].

1. INTRODUCTION
Since the Roman Empire, through the Inquisition and the Renaissance, until today, humanity has long debated the morality of warfare [4]. While it is universally acknowledged that peace is a preferable condition than war, this has not deterred the persistent conduct of lethal conflict over millennia. Referring to the improving technology of the day and its impact on the inevitability of warfare, in 1832 Clausewitz [5] stated "that the tendency to destroy the adversary which lies at the bottom of the conception of war is in no way changed or modified through the progress of civilization". More recently Cook [6] observed "The fact that constraints of just war are routinely overridden is no more a proof of their falsity and irrelevance than the existence of immoral behavior 'refutes' standards of morality: we know the standard, and we also know human beings fall short of that standard with depressing regularity".

St. Augustine is generally attributed, 1600 years ago, with laying the foundations of Christian Just War thought [6] and that Christianity helped humanize war by refraining from unnecessary killing [7]. Augustine (as reported via Aquinas) noted that emotion can clearly cloud judgment in warfare:

> *The passion for inflicting harm, the cruel thirst for vengeance, an unpacific and relentless spirit, the fever of revolt, the lust of power, and suchlike things, all these are rightly condemned in war* [4, p. 28].

Fortunately, these potential failings of man need not be replicated in autonomous battlefield robots.

From the 19th Century on, nations have struggled to create laws of war based on the principles of Just War Theory [7,8]. These laws speak to both *Jus in Bello*, which applies limitations to the conduct of warfare, and *Jus ad Bellum*, which restricts the conditions required prior to entering into war, where both form a major part of the logical underpinnings of the Just War tradition.

The advent of autonomous robotics in the battlefield, as with any new technology, is primarily concerned with *Jus in Bello*, i.e., defining what constitutes the ethical use of these systems during conflict, given military necessity. There are many questions that remain unanswered and even undebated within this context. At least two central principles are asserted from the Just War tradition: the principle of *discrimination* of military objectives and combatants from non-combatants and the structures of civil society; and the principle of *proportionality* of means, where acts of war should not yield damage disproportionate to the ends that justifies their use. Noncombatant harm is considered only justifiable when it is truly collateral, i.e., indirect and unintended, even if foreseen. Combatants retain certain rights as well, e.g.,

HRI'08, March 12–15, 2008, Amsterdam, Netherlands.

once they have surrendered and laid down their arms they assume the status of non-combatant and are no longer subject to attack. *Jus in Bello* also requires that agents must be held responsible for their actions in war [9]. This includes the consequences for obeying orders when they are known to be immoral as well as the status of ignorance in warfare. These aspects also need to be addressed in the application of lethality by autonomous systems, and as we will see in Section III, are hotly debated by philosophers.

The Laws of War (LOW), encoded in protocols such as the Geneva Conventions and Rules of Engagement (ROE), prescribe what is and what is not acceptable in the battlefield in both a global (Standing ROE) and local (Supplemental ROE) context, The ROE are required to be fully compliant with the laws of war. Defining these terms [10]:

- Laws of War – That part of international law that regulates the conduct of armed hostilities.
- Rules of Engagement - Directives issued by competent military authority that delineate the circumstances and limitations under which United States Forces will initiate and/or continue combat engagement with other forces encountered.

As early as 990, the Angiers Synod issued formal prohibitions regarding combatants' seizure of hostages and property [7]. The Codified Laws of War have developed over centuries, with Figure 1 illustrating several significant landmarks along the way.

Typical battlefield limitations, especially relevant with regard to the potential use of lethal autonomous systems, include [4,12]:

- Acceptance of surrender of combatants and the humane treatment of prisoners of war.
- Use of proportionality of force in a conflict.
- Protection of both combatants and non-combatants from unnecessary suffering.
- Avoiding unnecessary damage to property and people not involved in combat.
- Prohibition on attacking people or vehicles bearing the Red Cross or Red Crescent emblems, or those carrying a white flag and that are acting in a neutral manner.
- Avoidance of the use of torture on anyone for any reason.
- Non-use of certain weapons such as blinding lasers and small caliber high-velocity projectiles, in addition to weapons of mass destruction.

Waltzer sums it up: "... war is still, somehow, a rule-governed activity, a world of permissions and prohibitions – a moral world, therefore, in the midst of hell" [8, p. 36]. These laws of war continue to evolve over time as technology progresses, and any lethal autonomous system that attempts to adhere to them must similarly be able to adapt to new policies and regulations as they are formulated by international society.

Of course there are serious questions and concerns regarding the Just War tradition itself, often evoked by pacifists. Yoder [13] questions the premises on which it is built, and in so doing also raises some issues that potentially affect autonomous systems. For example he questions "Are soldiers when assigned a mission given sufficient information to determine whether this is an order they should obey? If a person under orders is convinced he or she must disobey, will the command structure, the society, and the church honor that dissent?" Clearly, if we embed an ethical "conscience" into an autonomous system, it is only as good as the information upon which it functions. It is a working assumption, perhaps naïve, that the autonomous agent ultimately will be provided with an amount of battlefield information equal to or greater than a human solider is capable of managing. This does seem a reasonable assumption, however, with the advent of network-centric warfare and the emergence of the Global Information Grid.

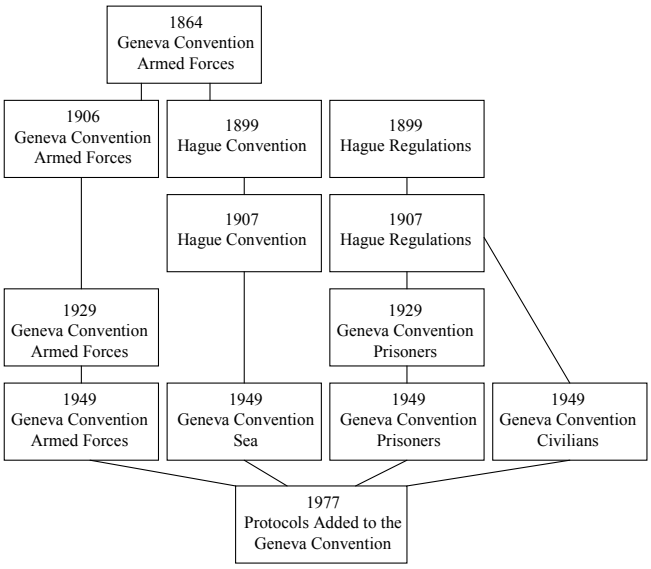

Fig 1: Development of Codified Laws of War (After [11])

It is also assumed in this work, that if an autonomous agent refuses to conduct an unethical action, it will be able to explain to some degree its underlying logic for such a refusal. If commanders were provided with the authority, by some means, to override the autonomous system's resistance to executing an order that it deems unethical, he or she in so doing would assume responsibility for the consequences of such an action.

These issues are but the tip of the iceberg of the ethical quandaries surrounding the deployment of autonomous systems capable of lethality. It is my contention, nonetheless, that if (or when) these systems are deployed in the battlefield, it is the roboticist's duty to ensure they are as safe as possible to both combatant and noncombatant alike, as is prescribed by our society's commitment to International Conventions encoded in the Laws of War, and other similar doctrine, e.g., the Code of Conduct and Rules of Engagement. The research in this article operates upon these underlying assumptions.

2. TRENDS TOWARDS LETHALITY IN THE BATTLEFIELD

There is only modest evidence that the application of lethality by autonomous systems is currently considered differently than any other weaponry. This is typified by informal commentary where some individuals state that a human will always be in the loop regarding the application of lethal force to an identified target. Often the use of the lethality in this context is considered more from a safety perspective [14], rather than a moral one. But if a

human being in the loop is the flashpoint of this debate, the real question is then at what level is the human in the loop? Will it be confirmation prior to the deployment of lethal force for each and every target engagement? Will it be at a high-level mission specification, such as "Take that position using whatever force is necessary"? Several military robotic automation systems already operate at the level where the human is in charge and responsible for the deployment of lethal force, but not in a directly supervisory manner. Examples include the Phalanx system for Aegis-class cruisers in the Navy, cruise missiles, or even (and generally considered as unethical due to their indiscriminate use of lethal force) anti-personnel mines or alternatively other more discriminating classes of mines, (e.g., anti-tank). These devices can even be considered to be robotic by some definitions, as they all are capable of sensing their environment and actuating, in these cases through the application of lethal force.

It is anticipated that teams of autonomous systems and human soldiers will work together on the battlefield, as opposed to the common science fiction vision of armies of unmanned systems operating by themselves. A range of unmanned robotic systems are already being developed or are already in use that employ lethal force such as the ARV (Armed Robotic Vehicle), a component of the Future Combat System (FCS); Predator UAVs (unmanned aerial vehicles) equipped with hellfire missiles that have already seen combat but under direct human supervision; and the development of an armed platform for use in the Korean Demilitarized Zone [15,16] to name a few. Some particulars follow:

- The South Korean robot platform mentioned above is intended to be able to detect and identify targets in daylight within a 4km radius, or at night using infrared sensors within a range of 2km, providing for either an autonomous lethal or non-lethal response. Although a designer of the system states that "the ultimate decision about shooting should be made by a human, not the robot", the system does have an automatic mode in which it is capable of making the decision on its own [17].
- iRobot, the maker of Roomba, is now providing versions of their Packbots capable of tasering enemy combatants [18]. This non-lethal response, however, does require a human-in-the-loop, unlike the South Korean robot under development.
- The SWORDS platform developed by Foster-Miller is already at work in Iraq and Afghanistan and is fully capable of carrying lethal weaponry (M240 or M249 machine guns, or a Barrett .50 Caliber rifle). [19]
- Israel is deploying stationary robotic gun-sensor platforms along its borders with Gaza in automated kill zones, equipped with fifty caliber machine guns and armored folding shields. Although it is currently only used in a remote controlled manner, an IDF division commander is quoted as saying "At least in the initial phases of deployment, we're going to have to keep a man in the loop", implying the potential for more autonomous operations in the future. [20]
- Lockheed-Martin, as part of its role in the Future Combat Systems program is developing an Armed Robotic Vehicle-Assault (Light) MULE robot weighing in at 2.5 tons. It will be armed with a line-of-sight gun and an anti-tank capability, to provide "immediate, heavy firepower to the dismounted soldier". [21]
- The U.S. Air Force has created their first hunter-killer UAV, named the MQ-9 Reaper. According to USAF General Moseley, the name Reaper is "fitting as it captures the lethal nature of this new weapon system". It has a 64 foot wingspan and carries 15 times the ordnance of the Predator, flying nearly three times the Predator's cruise speed. As of September 2006, 7 were already in inventory with more on the way. [22]
- The U.S. Navy for the first time is requesting funding for acquisition in 2010 of armed Firescout UAVs, a vertical-takeoff and landing tactical UAV that will be equipped with kinetic weapons. The system has already been tested with 2.75 inch unguided rockets. The UAVs are intended to deal with threats such as small swarming boats. As of this time the commander will determine whether or not a target should be struck. [23].

An even stronger indicator regarding the future role of autonomy and lethality appears in a recent U.S. Army Solicitation for Proposals [24], which states:

> *Armed UMS [Unmanned Systems] are beginning to be fielded in the current battlespace, and will be extremely common in the Future Force Battlespace… This will lead directly to the need for the systems to be able to operate autonomously for extended periods, and also to be able to collaboratively engage hostile targets within specified rules of engagement… with final decision on target engagement being left to the human operator****…. Fully autonomous engagement without human intervention should also be considered, under user-defined conditions, as should both lethal and non-lethal engagement and effects delivery means****.* [Boldface mine]

There is some evidence of restraint, however, in the use of unmanned systems designed for lethal operations, particularly regarding their autonomous use. A joint government industry council has generated a set of safety precepts [25] that bear this hallmark:

> DSP-6: The UMS [UnManned System] shall be designed to prevent uncommanded fire and/or release of weapons or propagation and/or radiation of hazardous energy.
>
> DSP-13: The UMS shall be designed to identify to the authorized entity(s) the weapon being released or fired.
>
> DSP-15: The firing of weapon systems shall require a minimum of two independent and unique validated messages in the proper sequence from authorized entity(ies), each of which shall be generated as a consequence of separate authorized entity action. Both messages should not originate within the UMS launching platform.

Nonetheless, the trend is clear: warfare will continue and autonomous robots will ultimately be deployed in its conduct. Given this, questions then arise regarding how these systems can conform as well or better than our soldiers with respect to adherence to the existing Laws of War. This paper focuses on this issue directly from a design perspective.

This is no simple task however. In the fog of war it is hard enough for a human to be able to effectively discriminate whether or not a target is legitimate. Fortunately for a variety of reasons, it may be

anticipated, despite the current state of the art, that in the future autonomous robots may be able to perform better than humans under these conditions, for the following reasons:

1. The ability to act conservatively: i.e., they do not need to protect themselves in cases of low certainty of target identification. UxVs do not need to have self-preservation as a foremost drive, if at all. They can be used in a self-sacrificing manner if needed and appropriate without reservation by a commanding officer.
2. The eventual development and use of a broad range of robotic sensors better equipped for battlefield observations than humans' currently possess.
3. They can be designed without emotions that cloud their judgment or result in anger and frustration with ongoing battlefield events. In addition, "Fear and hysteria are always latent in combat, often real, and they press us toward fearful measures and criminal behavior" [8, p. 251]. Autonomous agents need not suffer similarly.
4. Avoidance of the human psychological problem of "scenario fulfillment" is possible, a factor believed partly contributing to the downing of an Iranian Airliner by the USS Vincennes in 1988 [26]. This phenomena leads to distortion or neglect of contradictory information in stressful situations, where humans use new incoming information in ways that only fit their pre-existing belief patterns - a form of premature cognitive closure. Robots need not be vulnerable to such patterns of behavior.
5. They can integrate more information from more sources far faster than a human possibly could in real-time before responding with lethal force. This can arise from multiple remote sensors and intelligence (including human) sources, as part of the Army's network-centric warfare concept and the concurrent development of the Global Information Grid.
6. When working in a team of combined human soldiers and autonomous systems, they have the potential capability of independently and objectively monitoring ethical behavior in the battlefield by all parties and reporting infractions of human soldiers that might be observed. This presence might possibly lead to a reduction in human ethical infractions.

It is not my belief that an unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform more ethically that human soldiers are capable of. Unfortunately the trends in human behavior in the battlefield regarding their following legal and ethical requirements are questionable at best. A recent report from the Surgeon General's Office [27] regarding the battlefield ethics of soldiers and marines deployed in Operation Iraqi Freedom is disconcerting. The following findings are taken directly from that report:

1. Approximately 10% of Soldiers and Marines report mistreating non-combatants (damaged/destroyed Iraqi property when not necessary or hit/kicked a noncombatant when not necessary). Soldiers that have high levels of anger, experience high levels of combat or those who screened positive for a mental health problem were nearly twice as likely to mistreat non-combatants as those who had low levels of anger or combat or screened negative for a mental health problem.
2. Only 47% of Soldiers and 38% of Marines agreed that noncombatants should be treated with dignity and respect.
3. Well over a third of Soldiers and Marines reported torture should be allowed, whether to save the life of a fellow Soldier or Marine or to obtain important information about insurgents.
4. 17% of Soldiers and Marines agreed or strongly agreed that all noncombatants should be treated as insurgents.
5. Just under 10% of Soldiers and Marines reported that their unit modifies the ROE to accomplish the mission.
6. 45% of Soldiers and 60% of Marines did not agree that they would report a fellow soldier/marine if he had injured or killed an innocent noncombatant.
7. Only 43% of Soldiers and 30% of Marines agreed they would report a unit member for unnecessarily damaging or destroying private property.
8. Less than half of Soldiers and Marines would report a team member for an unethical behavior.
9. A third of Marines and over a quarter of Soldiers did not agree that their NCOs and Officers made it clear not to mistreat noncombatants.
10. Although they reported receiving ethical training, 28% of Soldiers and 31% of Marines reported facing ethical situations in which they did not know how to respond.
11. Soldiers and Marines are more likely to report engaging in the mistreatment of Iraqi noncombatants when they are angry, and are twice as likely to engage in unethical behavior in the battlefield than when they have low levels of anger.
12. Combat experience, particularly losing a team member, was related to an increase in ethical violations.

Possible explanations for the persistence of war crimes by combat troops are discussed in [28]. These include:

- High friendly losses leading to a tendency to seek revenge.
- High turnover in the chain of command, leading to weakened leadership.
- Dehumanization of the enemy through the use of derogatory names and epithets.
- Poorly trained or inexperienced troops.
- No clearly defined enemy.
- Unclear orders where intent of the order may be interpreted incorrectly as unlawful.

There is clearly room for improvement, and autonomous systems may help.

3. RELATED PHILOSOPHICAL THOUGHT

We now turn to several philosophers and practitioners who have specifically considered the military's potential use of lethal autonomous robotic agents. In a contrarian position regarding the use of battlefield robots, Sparrow [29] argues that any use of "fully autonomous" robots is unethical due to the *Jus in Bello* requirement that someone must be responsible for a possible war crime. His position is based upon deontological and consequentialist arguments. He argues that while responsibility could ultimately vest in the commanding officer for the system's

use, it would be unfair, and hence unjust, to both that individual and any resulting casualties in the event of a violation. Nonetheless, due to the increasing tempo of warfare, he shares my opinion that the eventual deployment of systems with ever increasing autonomy is inevitable. I agree that it is necessary that responsibility for the use of these systems must be made clear, but I do not agree that it is infeasible to do so. As mentioned earlier, several existing weapon systems are already in use that deploy lethal force autonomously to some degree, and they (with the exception of anti-personnel mines, due to their lack of discrimination, not responsibility attribution) are not considered generally to be unethical, at least to date.

Sparrow further draws parallels between robot warriors and child soldiers, both of which he claims cannot assume moral responsibility for their action. He neglects, however, to consider the possibility of the embedding of prescriptive ethical codes within the robot itself, which can govern its actions in a manner consistent with the Laws of War (LOW) and Rules of Engagement (ROE). This would seem to significantly weaken the claim he makes.

Along other lines, Sparrow [30], points out several clear challenges to the roboticist attempting to create a moral sense for a battlefield robot:

- "Controversy about right and wrong is endemic to ethics".
    - Response: While that is true, we have reasonable guidance by the agreed upon and negotiated Laws of War as well as the Rules of Engagement as a means to constrain behavior when compared to ungoverned solutions for autonomous robots.
- "I suspect that any decision structure that a robot is capable of instantiating is still likely to leave open the possibility that robots will act unethically."
    - Response: Agreed – It is the goal of this work to create systems that can perform better ethically than human soldiers do in the battlefield, albeit they will still be imperfect. This challenge seems achievable. Reaching perfection in almost anything in the real world, including human behavior, seems beyond our grasp.
- While he is "quite happy to allow that robots will become capable of increasingly sophisticated behavior in the future and perhaps even of distinguishing between war crimes and legitimate use of military force", the underlying question regarding responsibility, he contends, is not solvable (see also [29]).
    - Response: It is my belief by making the assignment of responsibility transparent and explicit, through the use of an architectural component serving as a responsibility advisor at all steps in the deployment of these systems, that this problem is indeed solvable.

Asaro [31] similarly argues from a position of loss of attribution of responsibility, but does broach the subject of robots possessing "moral intelligence". His definition of a moral agent seems applicable, where the agent adheres to a system of ethics, which it employs in choosing the actions that it either takes or refrains from taking. He also considers legal responsibility, which he states will compel roboticists to build ethical systems in the future. He notes, similar to what is proposed here, that if an existing set of ethical policies (e.g., LOW and ROE) is replicated through the robot's behavior, it enforces a particular morality in the robot itself. It is in this sense we strive to create such an ethical architectural component for unmanned systems, where the "particular morality" is derived from International Conventions concerning warfare.

One of the earliest arguments encountered based upon the difficulty to attribute responsibility and liability to autonomous agents in the battlefield was presaged by Perri 6 [32]. He assumes "at the very least the rules of engagement for the particular conflict have been programmed into the machines, and that only in certain types of emergencies are the machines expected to set aside these rules". I personally do not trust the view of setting aside the rules by the autonomous agent itself, as it begs the question of responsibility if it does so, but it may be possible for a human to assume responsibility for such a deviation if it is ever deemed appropriate (and ethical) to do so. The architecture proposed for this research [3] addresses specific issues regarding order refusal overrides by human commanders. While he rightly notes the inherent difficulty in attributing responsibility to the programmer, designer, soldier, commander, or politician for the potential of war crimes by these systems, it is believed that a deliberate and explicit assumption of responsibility by human agents for these systems can at least help focus such an assignment when required. An inherent part of the architecture for our project is a responsibility advisor [3], which specifically addresses these issues, although it would be naïve to say it will solve all of them. Case in point: often assigning and establishing responsibility for human war crimes, even through International Courts, is quite daunting.

Some would argue that the robot itself can be responsible for its own actions. Sullins [33], for example, is willing to attribute moral agency to robots far more easily than most, including myself, by asserting that simply if it is (1) in a position of responsibility relative to some other moral agent, (2) has a significant degree of autonomy, and (3) can exhibit some loose sort of intentional behavior ("there is no requirement that the actions really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction"), that it can then be considered to be a moral agent. Such an attribution unnecessarily complicates the issue of responsibility assignment for immoral actions. A perspective that a robot is incapable of becoming a moral agent that is fully responsible for its actions in any real sense, at least under present and near-term conditions, seems far more reasonable. Dennett [34] states that higher-order intentionality is a precondition for moral responsibility, including the opportunity for duplicity for example, something well beyond the capability of the sorts of robots under development in this article. Himma [35] requires that an artificial agent have both free will and deliberative capability before he is willing to attribute moral agency to it. Artificial (non-conscious) agents, in his view, have behavior that is either fully determined and explainable, or purely random in the sense of lacking causal antecedents. The bottom line for all of these lines of reasoning, at least for our purposes, is (and seemingly needless to say): for the sorts of autonomous agent architectures described in this paper, the robot is off the hook regarding responsibility. We will need to look toward humans for culpability for any ethical errors it makes in the lethal application of force.

But responsibility is not the lone sore spot for the potential use of

autonomous robots in the battlefield regarding Just War Theory. In a recent presentation Asaro [36] noted that the use of autonomous robots in warfare is unethical due to their potential lowering of the threshold of entry to war, which is in contradiction of *Jus ad Bellum*. One can argue, however, that this is not a particular issue limited to autonomous robots, but is typical for the advent of any significant technological advance in weaponry and tactics, and for that reason will not be considered in this paper. Other counterarguments could involve the resulting human-robot battlefield asymmetry as having a deterrent effect regarding entry into conflict by a State not in possession of the technology, which then might be more likely to sue for a negotiated settlement instead of entering into war. In addition, the potential for live or recorded data and video from gruesome real-time front-line conflict, possibly being made available to the media to reach into the living rooms of our nation's citizens, could lead to an even greater abhorrence of war by the general public rather than its acceptance[1]. Quite different imagery, one could imagine, as compared to the relatively antiseptic standoff precision high altitude bombings often seen in U.S. media outlets.

The Navy is examining the legal ramifications of the deployment of autonomous lethal systems in the battlefield [37], observing that a legal review is required of any new weapons system prior to their acquisition to ensure that it complies with the LOW and related treaties. To pass this review it must be shown that it does not act indiscriminately nor cause superfluous injury; in other words it must act with proportionality and discrimination, the hallmark criteria of *Jus in Bello*. The authors contend, and rightly so, that the problem of discrimination is the most difficult aspect of lethal unmanned systems, with only legitimate combatants and military objectives as just targets. They shift the paradigm for the robot to only identify and target weapons and weapon systems, not the individual(s) manning them. While they acknowledge several significant difficulties associated with this approach (e.g. spoofing and ruses to injure civilians), the question remains whether simply destroying weapons, without identifying those nearby as combatants and a lack of recognition of neighboring civilian objects, is legal in itself (i.e., ensuring that proportionality is exercised against a military objective). Nonetheless, it poses an interesting alternative where the system "targets the bow or arrow, not the archer". Their primary concerns arise from the acknowledged current limits on the ability to discriminate combatants from noncombatants. Although we are nowhere near providing robust methods to accomplish this in the near term, (except in certain limited circumstances with the use of friend-foe interrogation (FFI) technology), in my estimation considerable effort can and should be made into this research area by the DOD, and in many ways it already has, e.g., by using gait and other patterns of activity to identify suspicious persons. These very early steps, coupled with weapon recognition capabilities, could potentially provide even greater target discrimination than simply recognizing the weapons alone. Unique tactics (yet to be developed) for use by an unmanned system to actively ferret out the identity of a combatant by using a direct approach or other risk-taking (exposure) tactics can further illuminate what constitutes a legitimate target or not in the battlefield. This is an acceptable strategy by virtue of the robot's not needing to defend itself as a soldier would, perhaps by allowing the robot to employ self-sacrifice in order to reveal the presence of a combatant. There is no inherent right of self-defense for an autonomous system. In any case, clearly this is not a short-term research agenda.

The elimination of the need for an autonomous agent's claim of self-defense as an exculpation of responsibility through either justification or excuse is of related interest, which is a common occurrence during the occasioning of civilian casualties by human soldiers [38]. Robotic systems need make no appeal to self-defense or self-preservation in this regard, and can and should thus value civilian lives above their own continued existence. Of course there is no guarantee that a lethal autonomous system would be given that capability, but to be ethical I would contend that it must. This is a condition that a human soldier likely could not easily or ever attain to, and as such it would allow an ethical autonomous agent to potentially perform in a manner superior to that of a human in this regard. It should be noted that the system's use of lethal force does not preclude collateral damage to civilians and their property during the conduct of a military mission according to the Just War Principle of Double Effect[2], only that no claim of self-defense could be used to justify any such incidental deaths. It also does not negate the possibility of the autonomous system acting to defend fellow human soldiers under attack in the battlefield. Self-defense is not only for oneself, but also for one's fellow soldiers, army, and nation as defined in the LOW and Standing ROE.

Anderson [39], in his blog, points out the fundamental difficulty of assessing proportionality by a robot as required for *Jus in Bello*, largely due to the "apples and oranges" sorts of calculations that may be needed. He notes that a "practice", as opposed to a set of decision rules, will need to be developed, and although a daunting task, he sees it, in principle, as the same problem that humans have in making such a decision. Thus his argument is based on the degree of difficulty rather than any form of fundamental intransigence. Research in the area of combatant discrimination can provide the opportunity to make this form of reasoning regarding proportionality explicit. Indeed, different forms of reasoning beyond simple inference will be required, and case-based reasoning (CBR) is just one such candidate to be considered for use in the responsibility advisor [3]. We have already put CBR to work in intelligent robotic systems [40,41], where we reason from previous experience using analogy as appropriate. It may also be feasible to expand its use in the context of proportional use of force. Given the potential of a computer-based system to more effectively determine the consequences regarding the use of any particular weapon system on a given target faster than a human could (e.g., blast radius determination, etc.) via real-time simulation or estimation, it seems that selection of proportional means lies within the reach of autonomous lethal systems.

[1] BBC reporter Dan Damon pointed out this potential effect during an interview in July 2007.

[2] The Principle of Double Effect, derived from the Middle Ages, asserts "that while the death or injury of innocents is always wrong, either may be excused if it was not the intended result of a given act of war" [5, p.258]. As long as the collateral damage is an unintended effect (i.e., innocents are not deliberately targeted), it is excusable according to the LOW even if is foreseen (and that proportionality is adhered to).

Walzer comments on the issue of risk-free war-making, an imaginable outcome of the introduction of lethal autonomous systems. He states "there is no principle of Just War Theory that bars this kind of warfare" [42, p. 16]. Just War theorists have not discussed this issue to date and he states it is time to do so. Despite Walzer's assertion, discussions of this sort could possibly lead to prohibitions or restrictions on the use of lethal autonomous systems in the battlefield for this or any of the other reasons above. For example, Bring [43] states, for the more general case, "An increased use of standoff weapons is not to the advantage of civilians. The solution is not a prohibition of such weapons, but rather a reconsideration of the parameters for modern warfare as it affects civilians." Personally, I clearly support the start of such talks regarding the use of battlefield autonomous systems at any and all levels to clarify just what is acceptable internationally. In my view this warfighting proposition will not be risk-free, as teams of robots and soldiers will be working side-by-side in the battlefield, taking advantage of the principle of force multiplication where a single warfighter can project his presence as equivalent to several soldiers' capabilities of the past through the use of these systems. Substantial risk to human life will remain present (albeit significantly less so on the friendly side) in a clearly asymmetrical fashion.

I suppose a discussion of the ethical behavior of robots would be incomplete without some reference to Asimov's [44] "Three Laws of Robotics"[3] (there are actually four [45]). Needless to say, I am not alone in my belief that, while these Laws are elegant in their simplicity and have served a useful fictional purpose by bringing to light a whole range of issues surrounding robot ethics and rights, they are at best a strawman to bootstrap the ethical debate and as such serve no useful practical purpose beyond their fictional roots. Anderson [46], from a philosophical perspective, similarly rejects them, arguing: "Asimov's 'Three Laws of Robotics' are an unsatisfactory basis for Machine Ethics, regardless of the status of the machine". With all due respect, I must concur.

4. SUMMARY AND FUTURE WORK

This paper presents the background, motivation and philosophy for the design of an ethical autonomous robotic system capable of using lethal force. The system is governed by the Laws of War and Rules of Engagement using them as constraints. It is a goal of this research, which is funded by the Army Research Office, to yield ethical performance by autonomous systems that eventually exceed that of human soldiers.

Specific design models for the implementation of this approach appear in [3] which include an ethical governor, ethical behavioral control, an ethical adaptor, and a responsibility advisor. These are also discussed in subsequent papers in this series [1,2].

5. ACKNOWLEDGMENTS

This research is funded under Contract #W911NF-06-1-0252 from the U.S. Army Research Office.

[3] See http://en.wikipedia.org/wiki/Three_Laws_of_Robotics for a summary discussion of all 4 laws.

6. REFERENCES

[1] Arkin, R.C., "Governing Ethical Behavior: Embedding an Ethical Controller in a Hybrid Deliberative-Reactive Robot Architecture - Part II: Formalization for Ethical Control", *Proc. 1st Conference on Artificial General Intelligence*, Memphis, TN, March 2008.

[2] Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture - Part III: Representational and Architectural Considerations", *Proc. Technology in Wartime Conference*, Stanford, CA, Jan. 2008.

[3] Arkin, R.C., "Governing Ethical Behavior: Embedding an Ethical Controller in a Hybrid Deliberative-Reactive Robot Architecture", GVU Technical Report GIT-GVU-07-11, College of Computing, Georgia Tech, 2007.

[4] May, L., Rovie, E., and Viner, S., *The Morality of War: Classical and Contemporary Readings,* Pearson-Prentice Hall, 2005.

[5] Clausewitz, C. Von, "On the Art of War", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner 2005), Pearson-Prentice Hall, pp. 115-121, 1832.

[6] Cook, M., The Moral Warrior: Ethics and Service in the U.S. Military, State University of New York Press, 2004.

[7] Wells, D., (Ed.), *An Encyclopedia of War and Ethic*s, Greenwood Press, 1996.

[8] Walzer, M*., Just and Unjust Wars,* 4th Ed., Basic Books, 1977.

[9] Fieser, J. and Dowden, B., "Just War Theory", The Internet Encyclopedia of Philosophy, 2007

[10] Department of Defense Joint Publication 1-02, *Dictionary of Military and Associated Terms*, April 2001, Amended through June 2007.

[11] Hartle, A., *Moral Issues in Military Decision Making*, 2nd Ed., Revised, University Press of Kansas, 2004.

[12] Wikipedia, "Laws of War", 2007. http://en.wikipedia.org/wiki/Laws_of_war,.

[13] Yoder, J.H., "When War is Unjust: Being Honest in Just-War Thinking", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, et al, 2005), Pearson-Prentice Hall, pp. 153-159, 1984.

[14] Department of Defense, *Unmanned Systems Safety Guide for DOD Acquisition*, 1st Edition, Version .96, Jan. 2007.

[15] Argy, P., "Ethics Dilemma in Killer Bots", Australian IT News, June 14, 2007.

[16] Samsung Techwin, http://www.samsungtechwin.com/product/features/dep/SSystem_e/SSsystem.html, 2007.

[17] Kumagai, J., "A Robotic Sentry for Korea's Demilitarized Zone", IEEE Spectrum, March 2007.

[18] Jewell, M., "Taser, iRobot team up to arm robots", AP News Wire, June 2007.

[19] Foster-Miller Inc., "Products & Service: TALON Military Robots, EOD, SWORDS, and Hazmat Robots", http://www.foster-miller.com/lemming.htm, 2007.

[20] Opall-Rome, B., "Israel Wants Robotic Guns, Missiles to Guard Gaza", Defensenews.com, 2007.

[21] Lockheed-Martin, Mule /ARV-A(L), Fact Sheet, 2007.

[22] Air Force, "Reaper moniker given to MQ-9 Unmanned Aerial Vehicle", Official Website of the United States Air Force, 2006.

[23] Erwin, S., "For the First Time, Navy will Launch Weapons from Surveillance Drones", *National Defense*, June 2007.

[24] U.S. Army SBIR Solicitation 07.2, Topic A07-032 "Multi-Agent Based Small Unit Effects Planning and Collaborative Engagement with Unmanned Systems", pp. Army 57-68, 2007.

[25] Joint Government/Industry Unmanned Systems Safety Initiatives, "Programmatic / Design / Operational Safety Precepts Rev F", 2007.

[26] Sagan, S., "Rules of Engagement", in *Avoiding War: Problems of Crisis Management* (Ed. A. George), Westview Press, 1991.

[27] Surgeon General's Office, Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07, Final Report, Nov. 17, 2006.

[28] Bill, B. (Ed.), *Law of War Workshop Deskbook*, International and Operational Law Department, Judge Advocate General's School, June 2000.

[29] Sparrow, R., "Killer Robots*", Journal of Applied Philosophy*, Vol. 24, No.1, 2006.

[30] Sparrow, R., Personal Communication, July 2, 2007.

[31] Asaro, P., "What Should We Want From a Robot Ethic?" *International Review of Information Ethics*, Vol. 6, pp. 9-16, Dec. 2006.

[32] Perri 6, "Ethics, Regulation and the New Artificial Intelligence, Part II: Autonomy and Liability", *Information, Communication and Society*, 4:3, pp. 406-434, 2001.

[33] Sullins, J., "When is a Robot a Moral Agent?" *International Journal of information Ethics*, Vol. 6, 12, 2006.

[34] Dennett, D., "When HAL Kills, Who's to Blame?", in *HAL's Legacy: 2001's Computer as Dream and Reality*, (Ed. D. Stork), MIT Press, 1996.

[35] Himma, K., "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?" *7th International Computer Ethics Conference*, San Diego, CA, July 2007.

[36] Asaro, P., "How Just Could a Robot War Be?", presentation at *5th European Computing and Philosophy Conference*, Twente, NL, June 2007.

[37] Canning, J., Riggs, G., Holland, O., Blakelock, C., "A Concept for the Operation of Armed Autonomous Systems on the Battlefield", *Proc. AUVSI 2004*, Anaheim, CA, Aug. 2004.

[38] Woodruff, P., "Justification or Excuse: Saving Soldiers at the Expense of Civilians", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner, 2005), Pearson-Prentice Hall, pp. 281-291, 1982.

[39] Anderson, K., "The Ethics of Robot Soldiers?", *Kenneth Anderson's Law of Jaw and Just War Theory Blog*, July 4, 2007.

[40] Likhachev, M., Kaess, M., and Arkin, R.C., "Learning Behavioral Parameterization Using Spatio-Temporal Case-based Reasoning", *2002 IEEE International Conference on Robotics and Automation*, Washington, D.C., May 2002.

[41] Ram, A., Arkin, R.C., Moorman, K., and Clark, R.J., "Case-based Reactive Navigation: A case-based method for on-line selection and adaptation of reactive control parameters in autonomous robotic systems", *IEEE Transactions on Systems, Man, and Cybernetics*, Volume 27, Part B, No. 3, June 1997, pp. 376-394.

[42] Walzer, M., *Arguing About War*, Yale Univ. Press, 2004.

[43] Bring, O., "International Humanitarian Law After Kosovo: Is Lex Lata Sufficient?", in *Legal and Ethical Lessons of NATO's Kosovo Campaign*, International Law Studies (Ed. A. Wall), Naval War College, Vol. 78, pp. 257-272, 2002.

[44] Asimov, I. *I, Robot,* New York: Doubleday & Co., 1950.

[45] Asimov, I., *Robots and Empire*, New York: Doubleday & Company, 1985.

[46] Anderson, S., "Asimov's 'Three Laws of Robotics' and Machine Metaethics", *AI and Society*, Springer, published online March 2007.