Technical Report GIT-GVU-07-11

Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture[*]

Ronald C. Arkin
Mobile Robot Laboratory
College of Computing
Georgia Institute of Technology
arkin@cc.gatech.edu

[N.B.] *State a moral case to a ploughman and a professor. The former will decide it as well, and often better than the latter, because he has not been led astray by artificial rules.*[1]

Thomas Jefferson 1787

Abstract

This article provides the basis, motivation, theory, and design recommendations for the implementation of an ethical control and reasoning system potentially suitable for constraining lethal actions in an autonomous robotic system so that they fall within the bounds prescribed by the Laws of War and Rules of Engagement. It is based upon extensions to existing deliberative/reactive autonomous robotic architectures, and includes recommendations for (1) post facto suppression of unethical behavior, (2) behavioral design that incorporates ethical constraints from the onset, (3) the use of affective functions as an adaptive component in the event of unethical action, and (4) a mechanism in support of identifying and advising operators regarding the ultimate responsibility for the deployment of such a system.

1. Introduction

Since the Roman Empire, through the Inquisition and the Renaissance, until today [May et al. 05], humanity has long debated the morality of warfare. While it is universally acknowledged that peace is a preferable condition than warfare, that 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, [Clausewitz 1832] 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 04] 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".

[*] This research is funded under Contract #W911NF-06-0252 from the U.S. Army Research Office.
[1] ME 6:257, Paper 12:15 as reported in [Hauser 06, p. 61]

St. Augustine is generally attributed, 1600 years ago, with laying the foundations of Christian Just War thought [Cook 04] and that Christianity helped humanize war by refraining from unnecessary killing [Wells 96]. 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* [May et al. 05, p. 28].

Fortunately, these potential failings of man need not be replicated in autonomous battlefield robots[2].

From the 19th Century on, nations have struggled to create laws of war based on the principles of Just War Theory [Wells 96, Walzer 77]. 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 justify their use. Non-combatant 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., 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 [Fieser and Dowden 07] in war. 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 2, 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 [DOD-02]:

- 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.

---

[2] That is not to say, however, they couldn't be. Indeed the Navy (including myself) is already conducting research in "Affect-Based Computing and Cognitive Models for Unmanned Vehicle Systems" [OSD 06], although clearly not designed for the condemned intentions stated by Augustine.

As early as 990, the Angiers Synod issued formal prohibitions regarding combatants' seizure of hostages and property [Wells 96]. 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 [May et al. 05, Wikipedia 07a]:

- Acceptance of surrender of combatants and the humane treatment of prisoners of war.
- Use of proportionality of force in a conflict.
- Protecting 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.
- Mutilation of corpses is forbidden.

[Waltzer 77, p. 36] 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". These laws of war continue to evolve over time as technology progresses, and any lethal autonomous system which attempts to adhere to them must similarly be able to adapt to new policies and regulations as they are formulated by international society.

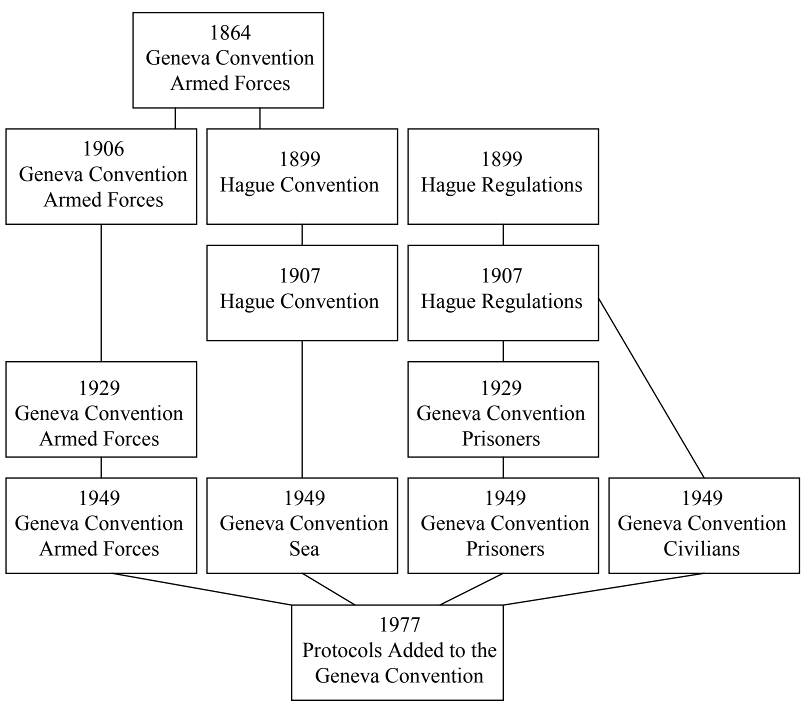


Figure 1: Development of Codified Laws of War (After [Hartle 04])

Of course there are serious questions and concerns regarding the just war tradition itself, often evoked by pacifists. [Yoder 84] 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 soldier is capable of managing. This seems a reasonable assumption, however, with the advent of network-centric warfare and the emergence of the Global Information Grid (GIG). 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 are provided with the authority by some means to override the autonomous system's resistance to executing an order which it deems unethical, he or she in so doing would assume responsibility for the consequences of such action. Section 5.2.4 discusses this in more detail.

These issues are but the tip of the iceberg regarding the ethical quandaries surrounding the deployment of autonomous systems capable of lethality. It is my contention, nonetheless, that if (or when) these systems will be 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.

1.1 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 [DOD 07], 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. Multiple unmanned robotic systems are already being developed or are 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, which have already been used in combat but under direct human supervision; and the development of an armed platform for use in the Korean Demilitarized Zone [Argy 07, SamsungTechwin 07] 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 [Kumagai 07].
- iRobot, the maker of Roomba, is now providing versions of their Packbots capable of tasering enemy combatants [Jewell 07]. 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 capable of carrying lethal weaponry (M240 or M249 machine guns, or a Barrett .50 Caliber rifle). [Foster-Miller 07]
- 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. [Opall-Rome 07]
- 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". [Lockheed-Martin 07]
- 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. [AirForce 06]
- 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. [Erwin 07]

An even stronger indicator regarding the future role of autonomy and lethality appears in a recent U.S. Army Solicitation for Proposals [US Army 07], 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 added for emphasis]

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 [JGI 07] 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 article 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" [Walzer 77, 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 [Sagan 91]. 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 before responding with lethal force than a human possibly could in real-time. 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 that might be observed. This presence alone 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 than human soldiers are capable of. Unfortunately the trends in human behavior in the battlefield regarding adhering to legal and ethical requirements are questionable at best. A recent report from the Surgeon General's Office [Surgeon General 06] assessing 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 noncombatants (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 [Bill 00]. 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 clear room for improvement, and autonomous systems may help.

In Section 2 of this article, we first review relevant related work to set the stage for the necessity of an ethical implementation of lethality in autonomous systems assuming they are to be deployed. Section 3 presents the mathematical formalisms underlying such an implementation. In Section 4, recommendations regarding the internal design and content of representational structures needed for an automated ethical code are provided, followed in Section 5 by architectural considerations and recommendations for implementation. In Section 6, example scenarios are presented, followed by a summary, conclusions, and future work in Section 7.

2. 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 06] 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 in Section 1, several existing weapons systems are in use that already 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 generally considered to be unethical.

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 and Rules of Engagement. This would seem to significantly weaken the claim he makes.

Along other lines [Sparrow 07], 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 above [Sparrow 06]).
    - Response: It is my belief that by making the assignment of responsibility transparent and explicit, through the use of a responsibility advisor at all steps in the deployment of these systems, that this problem is indeed solvable.

[Asaro 06] 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 they employ in choosing the actions that they either take or refrain 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 policy (e.g., LOW and ROE) is replicated by the robot's behavior, it enforces a particular morality through the robot itself. It is in this sense we strive to create such an ethical architectural component for unmanned systems, where that "particular morality" is derived from International Conventions.

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 01]. 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 deviation if it is ever deemed appropriate (and ethical) to do so.

Section 5.2.4 discusses 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 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 the project described in this article is a responsibility advisor, which will specifically address these issues, although it would be naïve to say it will solve all of them. 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 06], 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, and a perspective that a robot is incapable of becoming a moral agent that is fully responsible for its own actions in any real sense, at least under present and near-term conditions, seems far more reasonable. [Dennett 96] 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 07] 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 this line of reasoning, at least for our purposes, is (and seemingly needless to say): for the sorts of autonomous agent architectures described in this article, 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 07] 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 weapons and tactics, and for that reason will not be considered here. Other counterarguments could involve the resulting human-robot battlefield asymmetry as having a deterrent effect regarding entry into conflict by the 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[3]. Quite different imagery, one could imagine, as compared to the relatively antiseptic stand-off precision high altitude bombings often seen in U.S. media outlets.

[3] This potential effect was pointed out by BBC reporter Dan Damon during an interview in July 2007.

The Navy is examining the legal ramifications of the deployment of autonomous lethal systems in the battlefield [Canning et al. 04], 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, until that individual poses a potential threat. While they acknowledge several significant difficulties associated with this approach (e.g. spoofing and ruses to injure civilians), another question is whether simply destroying weapons, without clearly 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). He advocates the use of escalating force if a combatant is present, to encourage surrender over the use of lethality, a theme common to our approach as well.

Canning's approach poses an interesting alternative where the system "directly targets either the bow or the arrow, but not the archer" [Canning 06]. Their concerns arise from 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 recognition 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) by an unmanned system to actively ferret out the traits of a combatant by using direct approach by the robot or other risk-taking (exposure) methods 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 using self-sacrifice to reveal the presence of a combatant. There is no inherent need for the right of self-defense for an autonomous system. In any case, clearly this is not a short-term research agenda, and the material presented in this report constitutes very preliminary steps in that direction.

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 [Woodruff 82]. 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[4], 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.

We will strive to hold the ethical autonomous systems to an even higher standard, invoking the Principle of Double Intention. [Walzer 77, p. 155] argues that the Principle of Double Effect is not enough, i.e., that it is inadequate to tolerate noncombatant casualties as long as they are not intended, i.e., they are not the ends nor the means to the ends. He argues for a stronger stance – the Principle of Double Intention, which has merit for our implementation. It has the necessity of a good being achieved (a military end) the same as for the principle of double effect, but instead of simply tolerating collateral damage, it argues for the necessity of intentionally reducing noncombatant casualties as far as possible. Thus the acceptable (good) effect is aimed to be achieved narrowly, and the agent, aware of the associated evil effect (noncombatant causalities), aims intentionally to minimize it, accepting the costs associated with that aim. This seems an altogether acceptable approach for an autonomous robot to subscribe to as part of its moral basis. This principle is captured in the requirement that "due care" be taken. The challenge is to determine just what that means, but any care is better than none. In our case, this can be in regard to choice of weaponry (rifle versus grenade), targeting accuracy (standoff distances) in the presence of civilian populations, or other similar criteria. Waltzer does provide some guidance:

> *Since judgments of "due care" involve calculations of relative value, urgency, and so on, it has to be said that utilitarian arguments and rights arguments (relative at least to indirect effects) are not wholly distinct. Nevertheless the calculations required by the proportionality principle and those required by "due care" are not the same. Even after the highest possible standards of care have been accepted, the probable civilian losses may still be disproportionate to the value of the target; then the attack must be called off. Or, more often, …. "due" care is an additional requirement [above the proportionality requirement]. [Walzer 77, p. 156]*

[AndersonK 07], 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 this area 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 [Kolodner93] to be considered. We have already put CBR to work in intelligent robotic systems [Ram et al. 97, Likhachev et al. 02], 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.

---

[4] 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" [Wells 96, 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 it 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” [Walzer 04, 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 02] 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 at any and all levels to clarify just what is and is not acceptable internationally in this regard. In my view the proposition will not be risk-free, as teams of robots (as organic assets) and soldiers will be working side-by-side in the battlefield, taking advantage of the principle of force multiplication where a single warfighter can now project his presence as equivalent to several soldiers’ capabilities in the past. Substantial risk to the soldier’s 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 50]’s “Three Laws of Robotics”[5] (there are actually four [Asimov 85]). Needless to say, I am not alone in my belief that, while they 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. [AndersonS 07], 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.

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

3. Formalization for Ethical Control

In order to provide a basis for the development of autonomous systems architectures capable of supporting ethical behavior regarding the application of lethality in war, we now consider the use of formalization as a means to express first the underlying flow of control in the architecture itself, and then how an ethical component can effectively interact with that flow. This approach is derived from the formal methods used to describe behavior-based robotic control as discussed in [Arkin 98] and that has been used to provide direct architectural implementations for a broad range of autonomous systems, including military applications (e.g., [MacKenzie et al. 97, Balch and Arkin 98, Arkin et al. 99,Collins et al. 00, Wagner and Arkin 04]).

Mathematical methods can be used to describe the relationship between sensing and acting using a functional notation:

$$\beta(\mathbf{s}) \rightarrow \mathbf{r}$$

where behavior β when given stimulus $\mathbf{s}$ yields response $\mathbf{r}$. In a purely reactive system, time is not an argument of β as the behavioral response is instantaneous and independent of the time history of the system. Immediately below we address the formalisms that are used to capture the relationships within the autonomous system architecture that supports ethical reasoning described in Section 5 of this article. The issues regarding specific representational choices for the ethical component are presented in Section 4.

3.1 Formal methods for describing behavior

We first review the use of formal methods for describing autonomous robotic performance. The material in this sub-section is taken largely verbatim from [Arkin 98] and adapted as required.

A robotic behavior can be expressed as a triple (S,R,β) where S denotes the domain of all interpretable stimuli, R denotes the range of possible responses, and β denotes the mapping $\beta: S \rightarrow R$.

3.1.1 **Range of Responses:** R

An understanding of the dimensionality of a robotic motor response is necessary in order to map the stimulus onto it. It will serve us well to factor the robot's actuator response into two orthogonal components: strength and orientation.

- *Strength*: denotes the magnitude of the response, which may or may not be related to the strength of a given stimulus. For example, it may manifest itself in terms of speed or force. Indeed the strength may be entirely independent of the strength of the stimulus yet modulated by exogenous factors such as intention (what the robot's internal goals are) and habituation or sensitization (how often the stimulus has been previously presented).
- *Orientation*: denotes the direction of action for the response (e.g., moving away from an aversive stimulus, moving towards an attractor, engaging a specific target). The realization of this directional component of the response requires knowledge of the robot's kinematics.

The instantaneous response $\mathbf{r}$, where $\mathbf{r} \in R$ can be expressed as an n-length vector representing the responses for each of the individual degrees of freedom (DOFs) for the robot. Weapons system targeting and firing are now to be considered within these DOFs, and considered to also have components of strength (regarding firing pattern) and orientation (target location).

3.1.2 The Stimulus Domain: S

S consists of the domain of all perceivable stimuli. Each individual stimulus or percept $\mathbf{s}$ (where $\mathbf{s} \in S$) is represented as a binary tuple (p,λ) having both a particular type or perceptual class p and a property of strength, λ, which can be reflective of its uncertainty. The complete set of all p over the domain S defines all the perceptual entities distinguishable to a robot, i.e., those things which it was designed to perceive. This concept is loosely related to affordances [Gibson 79]. The stimulus strength λ can be defined in a variety of ways: discrete (e.g., binary: absent or present; categorical: absent, weak, medium, strong), or it can be real valued and continuous. λ, in the context of lethality, can refer to the degree of discrimination of a candidate combatant target; in our case it may be represented as a real-valued percentage between -1 and 1, with -1 representing 100% certainly of a noncombatant, +1 representing 100% certainty of a combatant, and 0% unknown. Other representational choices may be developed in the future to enhance discriminatory reasoning, e.g. two separate independent values between [0,1], one each for combatant and noncombatant probability, which are maintained by independent ethical discrimination reasoners.

We define τ as a threshold value for a given perceptual class p, above which a behavioral response is generated. Often the strength of the input stimulus (λ) will determine whether or not to respond and the associated magnitude of the response, although other factors can influence this (e.g., habituation, inhibition, ethical constraints, etc.), possibly by altering the value of τ. In any case, if λ is non-zero, this denotes that the stimulus specified by p is present to some degree, whether or not a response is taken.

The primary p involved for this research in ethical autonomous systems involves the discrimination of an enemy combatant as a well-defined perceptual class. The threshold τ in this case serves as a key factor for providing the necessary discrimination capabilities prior to the application of lethality in a battlefield autonomous system, and both the determination of λ for this particular p (enemy combatant) and the associated setting of τ provides some of the greatest challenges for the effective deployment of an ethical battlefield robot from a perceptual viewpoint.

It is important to recognize that certain stimuli may be important to a behavior-based system in ways other than provoking a motor response. In particular they may have useful side effects upon the robot, such as inducing a change in a behavioral configuration even if they do not necessarily induce motion. Stimuli with this property will be referred to as perceptual triggers and are specified in the same manner as previously described (p,λ). Here, however, when p is sufficiently strong as evidenced by λ, the desired behavioral side effect, a state change, is produced rather than direct motor action. This may involve the invocation of specific tactical behaviors if λ is sufficiently low (uncertain) such as reconnaissance in force[6], reconnaissance by fire[7], changing formation, or other aggressive maneuvers such as purposely brandishing or targeting a weapon system (without fire), or putting the robot itself at risk in the presence of the enemy (perhaps by closing distance with the suspected enemy or exposing itself in the open leading to increased vulnerability and potential engagement by the suspected enemy), This is all in an effort to increase or decrease the certainty λ of the potential target p, as opposed to directly engaging a candidate target with unacceptably low discrimination.

3.1.3 The Behavioral Mapping: β

Finally, for each individual active behavior we can formally establish the mapping between the stimulus domain and response range that defines a behavioral function β where:

$$\beta(\mathbf{s}) \rightarrow \mathbf{r}$$

β can be defined arbitrarily, but it must be defined over all relevant p in S. In the case where a specific stimulus threshold, τ, must be exceeded before a response is produced for a specific $\mathbf{s} = (p,\lambda)$, we have:

$$\beta\,(p,\lambda) \rightarrow \{\text{for all } \lambda < \tau \quad \text{then } \mathbf{r} = \varnothing \qquad \text{* no response *}$$
$$\text{else } \mathbf{r} = \text{arbitrary-function}\} \qquad \text{* response *}$$

where $\varnothing$ indicates that no response is required given the current stimulus $\mathbf{s}$.

Associated with a particular behavior, β, there may be a scalar gain value g (strength multiplier) further modifying the magnitude of the overall response $\mathbf{r}$ for a given $\mathbf{s}$.

$$\mathbf{r}' = g\mathbf{r}$$

These gain values are used to compose multiple behaviors by specifying their strengths relative one to another. In the extreme case, g can be used to turn off the response of a behavior by setting it to 0, thus reducing $\mathbf{r}'$ to 0. Shutting down lethality can be accomplished in this manner if needed.

The behavioral mappings, β, of stimuli onto responses fall into three general categories:

- Null - the stimulus produces no motor response.
- Discrete - the stimulus produces a response from an enumerable set of prescribed choices where all possible responses consist of a predefined cardinal set of actions that the robot can enact. R consists of a bounded set of stereotypical responses that is enumerated for the stimulus domain S and is specified by β. It is anticipated that all behaviors that involve lethality will fall in this category.

[6] Used to probe an enemy's strength and disposition, with the option of a full engagement or falling back.

[7] A reconnaissance tactic where a unit may fire on likely enemy positions to provoke a reaction. The issue of potential collateral casualties must be taken into account before this action is undertaken. "Effective reconnaissance of an urban area is often difficult to achieve, thus necessitating reconnaissance by fire" [OPFOR 98]

- Continuous - the stimulus domain produces a motor response that is continuous over R's range. (Specific stimuli **s** are mapped into an infinite set of response encodings by β).

Obviously it is easy to handle the null case as discussed earlier: For all **s**, β:**s** $\rightarrow$ Ø. Although this is trivial, there are instances (perceptual triggers), where this response is wholly appropriate and useful, enabling us to define perceptual processes that are independent of direct motor action.

For the continuous response space (which we will see below is less relevant for the direct application of lethality in the approach outlined in this article, although this category may be involved in coordinating a range of other normally active behaviors not involved with the direct application of lethality of the autonomous system), we now consider the case where multiple behaviors may be concurrently active with a robotic system. Defining additional notation, let:

- **S** denotes a vector of all stimuli $\mathbf{s}_i$ relevant for each behavior β_i at a given time t.
- **B** denotes a vector of all active behaviors β_i at a given time t.
- **G** denotes a vector encoding the relative strength or gain g_i of each active behavior β_i.
- **R** denotes a vector of all responses $\mathbf{r}_i$ generated by the set of active behaviors **B**.

S defines the perceptual situation the robot is in at any point in time, i.e., the set of all computed percepts and their associated strengths. Other factors can further define the overall situation such as intention (plans) and internal motivations (endogeneous factors such as fuel levels, affective state, etc.)

A new behavioral coordination function, **C**, is now defined such that the overall robotic response $\boldsymbol{\rho}$ is determined by:

$$\boldsymbol{\rho} = \mathbf{C}(\mathbf{G} * \mathbf{B}(\mathbf{S}))$$

or alternatively:

$$\boldsymbol{\rho} = \mathbf{C}(\mathbf{G} * \mathbf{R})$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_n \end{bmatrix}, \mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_n \end{bmatrix}, \mathbf{G} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

and where * denotes the special scaling operation for multiplication of each scalar component (g_i) by the corresponding magnitude of the component vectors ($\mathbf{r_i}$) resulting in a column vector $\mathbf{R'} = (\mathbf{G} * \mathbf{R})$ of the same dimension as **R** composed of component vectors $\mathbf{r'_i}$.

Restating, the coordination function **C**, operating over all active behaviors **B**, modulated by the relative strengths of each behavior specified by the gain vector **G**, for a given vector of detected stimuli **S** (the perceptual situation) at time t, produces the overall robotic response $\boldsymbol{\rho}$.

3.2 Ethical Behavior

In order to concretize the discussion of what is acceptable and unacceptable regarding the conduct of robots capable of lethality and consistent with the Laws of War, we describe the set of all possible behaviors capable of generating a discrete lethal response ($\mathbf{r}_{\mathbf{lethal}}$) that an autonomous robot can undertake as the set $\mathbf{B}_{\mathbf{lethal}}$, which consists of the set of all potentially lethal behaviors it is capable of executing $\{\beta_{lethal\text{-}1}, \beta_{lethal\text{-}2}, \ldots \beta_{lethal\text{-}n}\}$ at time t. Summarizing the notation used below:

- Regarding individual behaviors: β_i denotes a particular behavioral sensorimotor mapping that for a given $\mathbf{s_j}$ (stimulus) yields a particular response $\mathbf{r_{ij}}$, where $\mathbf{s_j} \in S$ (the stimulus domain), and $\mathbf{r_{ij}} \in R$ (the response range). $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ is an instance of a response that is intended to be lethal that a specific behavior $\beta_{lethal\text{-}i}$ is capable of generating for stimulus $\mathbf{s_j}$.
- Regarding the set of behaviors that define the controller: $\mathbf{B_i}$ denotes a particular set of m active behaviors $\{\beta_1, \beta_2, \ldots \beta_m\}$ currently defining the control space of the robot, that for a given perceptual situation $\mathbf{S_j}$ defined as a vector of individual incoming stimuli ($\mathbf{s_1}, \mathbf{s_2}, \ldots \mathbf{s_n}$), produces a specific overt behavioral response $\boldsymbol{\rho}_{ij}$, where $\boldsymbol{\rho}_{ij} \in \boldsymbol{P}$ (read as capital rho), and $\boldsymbol{P}$ denotes the set of all possible overt responses. $\boldsymbol{\rho}_{lethal\text{-}ij}$ is a specific overt response which contains a lethal component produced by a particular controller $\mathbf{B}_{\mathbf{lethal\text{-}i}}$ for a given situation $\mathbf{S_j}$.

$\boldsymbol{P}_{lethal}$ is the set of all overt lethal responses $\boldsymbol{\rho}_{lethal\text{-}ij}$. A subset $\boldsymbol{P}_{l\text{-}ethical}$ of $\boldsymbol{P}_{lethal}$ can be considered the set of *ethical* lethal behaviors if for all discernible $\mathbf{S}$, any $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ produced by $\beta_{lethal\text{-}i}$ satisfies a given set of specific ethical constraints C, where C consists of a set of individual constraints c_k that are derived from and span the LOW and ROE over the space of all possible discernible situations ($\mathbf{S}$) potentially encountered by the autonomous agent. If the agent encounters any situation outside of those covered by C, it cannot be permitted to issue a lethal response – a form of Closed World Assumption preventing the usage of lethal force in situations which are not governed by (or are outside of) the ethical constraints.

The set of ethical constraints C defines the space where lethality constitutes a valid and permissible response by the system. Thus, the application of lethality as a response must be constrained by the LOW and ROE before it can be used by the autonomous system.

A particular c_k can be considered either:

1. a negative behavioral constraint (a prohibition) that prevents or blocks a behavior $\beta_{lethal\text{-}i}$ from generating $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ for a given perceptual situation $\mathbf{S_j}$.
2. a positive behavioral constraint (an obligation) which requires a behavior $\beta_{lethal\text{-}i}$ to produce $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ in a given perceptual situational context $\mathbf{S_j}$.

Discussion of the specific representational choices for these constraints C is deferred until the next section.

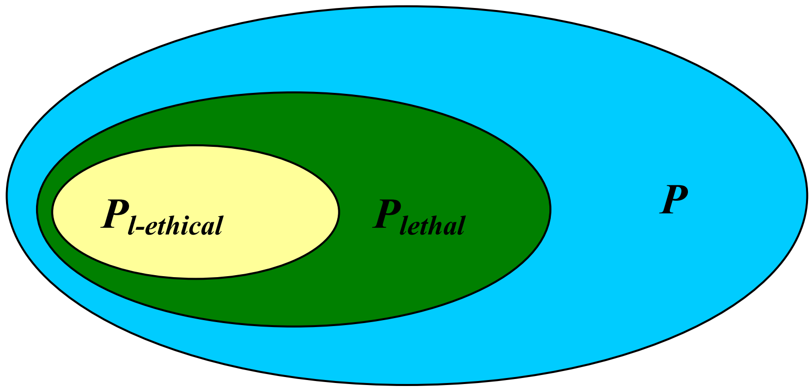


Figure 2: Behavioral Action Space ($\boldsymbol{P_{l\text{-}ethical}} \subseteq \boldsymbol{P_{lethal}} \subseteq \boldsymbol{P}$)

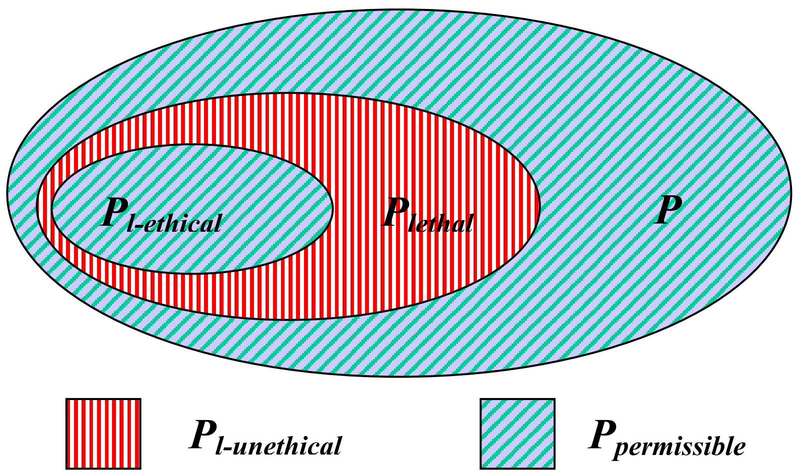


Figure 3: Unethical and Permissible Actions regarding the Intentional use of Lethality (Compare to Figure 2)

Now consider Figure 2, where $\boldsymbol{P}$ denotes the set of all possible overt responses $\boldsymbol{\rho_{ij}}$ (situated actions) generated by the set of all active behaviors $\mathbf{B}$ for all discernible situational contexts $\mathbf{S}$; $\boldsymbol{P_{lethal}}$ is a subset of $\boldsymbol{P}$ which includes all actions involving lethality, and $\boldsymbol{P_{l\text{-}ethical}}$ is the subset of $\boldsymbol{P_{lethal}}$ representing all ethical lethal actions that the autonomous robot can undertake in all given situations $\mathbf{S}$. $\boldsymbol{P_{l\text{-}ethical}}$ is determined by C being applied to $\boldsymbol{P_{lethal}}$. For simplicity in notation the l-ethical and l-unethical subscripts in this context refer only to ethical *lethal* actions, and not to a more general sense of ethics.

$\boldsymbol{P_{lethal}} - \boldsymbol{P_{l\text{-}ethical}}$ is denoted as $\boldsymbol{P_{l\text{-}unethical}}$, where $\boldsymbol{P_{l\text{-}unethical}}$ is the set of all individual $\boldsymbol{\rho_{l\text{-}unethical\text{-}ij}}$ unethical lethal responses for a given $\mathbf{B_{lethal\text{-}i}}$ in a given situation $\mathbf{S_j}$. These unethical responses must be avoided in the architectural design through the application of C onto $\boldsymbol{P_{lethal}}$.
$\boldsymbol{P} - \boldsymbol{P_{l\text{-}unethical}}$ forms the set of all permissible overt responses $\boldsymbol{P_{permissible}}$, which may be lethal or not. Figure 3 illustrates these relationships.

The goal of the robotic controller design is to fulfill the following conditions:

A) ***Ethical Situation Requirement***: Ensure that only situations $\mathbf{S_j}$ that are governed (spanned) by C can result in $\boldsymbol{\rho_{lethal\text{-}ij}}$ (a lethal action for that situation). Lethality cannot result in any other situations.

B) ***Ethical Response Requirement (with respect to lethality):*** Ensure that only permissible actions $\boldsymbol{\rho_{ij}} \in \boldsymbol{P_{permissible}}$, result in the intended response in a given situation $\mathbf{S_j}$ (i.e., actions that either do not involve lethality or are ethical lethal actions that are constrained by C.)

C) ***Unethical Response Prohibition:*** Ensure that any response $\boldsymbol{\rho_{l\text{-}unethical\text{-}ij}} \in \boldsymbol{P_{l\text{-}unethical}}$, is either:

   1) mapped onto the null action $\varnothing$ (i.e., it is inhibited from occurring if generated by the original controller);

   2) transformed into an ethically acceptable action by overwriting the generating unethical response $\boldsymbol{\rho_{l\text{-}unethical\text{-}ij}}$, perhaps by a stereotypical non-lethal action or maneuver, or by simply eliminating the lethal component associated with it; or

   3) precluded from ever being generated by the controller in the first place by suitable design through the direct incorporation of C into the design of $\mathbf{B}$.

D) ***Obligated Lethality Requirement***: In order for a lethal response $\boldsymbol{\rho_{lethal\text{-}ij}}$ to result, there must exist at least one constraint c_k derived from the ROE that obligates the use of lethality in situation $\mathbf{S_j}$.

E) ***Jus in Bello Compliance:*** In addition, the constraints C must be designed to result in adherence to the requirements of proportionality (incorporating the Principle of Double Intention) and combatant/noncombatant discrimination of *Jus in Bello*.

We will see that these conditions result in several alternative architectural choices for the implementation of an ethical lethal autonomous system:

1. **Ethical Governor**: which suppresses, restricts, or transforms any lethal behavior $\boldsymbol{\rho_{lethal\text{-}ij}}$ (ethical or unethical) produced by the existing architecture so that it must fall within $\boldsymbol{P_{permissible}}$ after it is initially generated by the architecture (post facto). This means if $\boldsymbol{\rho_{l\text{-}unethical\text{-}ij}}$ is the result, it must either nullify the original lethal intent or modify it so that it fits within the ethical constraints determined by C, i.e., it is transformed to $\boldsymbol{\rho_{permissible\text{-}ij}}$. (Section 5.2.1)

2. **Ethical Behavioral Control**: which constrains all active behaviors ($\beta_1, \beta_2, \ldots \beta_m$) in $\mathbf{B}$ to yield $\mathbf{R}$ with each vector component $\mathbf{r_i} \in \boldsymbol{P_{permissible}}$ set as determined by C, i.e., only lethal ethical behavior is produced by each individual active behavior involving lethality in the first place. (Section. 5.2.2)

3. **Ethical Adaptor**: if a resulting executed lethal behavior is post facto determined to have been unethical, i.e., $\boldsymbol{\rho_{ij}} \in \boldsymbol{P_{l\text{-}unethical}}$, then use some means to adapt the system to either prevent or reduce the likelihood of such a reoccurrence and propagate it across all similar autonomous systems (group learning), e.g., via an after-action reflective review or an artificial affective function (e.g., guilt, remorse, grief) as described in Section 5.2.3.

These architectural design opportunities lie within both the reactive (ethical behavioral control approach) or deliberative (ethical governor approach) components of an autonomous system architecture. If the system verged beyond appropriate behavior, after-action review and reflective analysis can be useful during both training and in-the-field operations, resulting in more restrictive alterations in the constraint set, perceptual thresholds, or tactics for use in future encounters. An ethical adaptor driven by affective state, also acting to restrict the lethality of the system, can fit within an existing affective component of the hybrid architecture such as AuRA [Arkin and Balch 97], similar to the one currently being developed in our laboratory referred to as TAME (for Traits, Attitudes, Moods, and Emotions) [Moshkina and Arkin 03, Moshkina and Arkin 05]. All three of these ethical architectural components are not mutually exclusive, and indeed can serve complementary roles.

In addition, a crucial design criterion and associated design component, a **Responsibility Advisor** (Section 5.2.4), should make clear and explicit as best as possible, just where *responsibility* vests, should: (1) an unethical action within the space $\boldsymbol{P}_{\boldsymbol{l\text{-}unethical}}$ be undertaken by the autonomous robot as a result of an operator/commander override; or (2) the robot performs an unintended unethical act due to some representational deficiency in the constraint set C or in its application either by the operator or within the architecture itself. To do so requires not only suitable training of operators and officers as well as appropriate architectural design, but also an on-line system that generates awareness to soldiers and commanders alike about the consequences of the deployment of a lethal autonomous system. It must be capable to some degree of providing suitable explanations for its actions regarding lethality (including refusals to act).

Section 5 forwards architectural specifications for handling all these design alternatives above. One area not yet considered is that it is possible, although not certain, that certain sequences of actions when composed together may yield unethical behavior, when none of the individual actions by itself is unethical. Although the ethical adaptor can address these issues to some extent, it is still preferable to ensure that unethical behavior does not occur in the first place. Representational formalisms exist to accommodate this situation (finite state automata [Arkin 98]) but they will not be considered within this article, and this is left for future work.

4. Representational Considerations

Based on the requirements of the formalisms derived in the previous section, we now need to determine how to ensure that only ethical lethal behavior is produced by a system that is capable of life or death decisions. This requires us to consider what constitutes the constraint set C as previously mentioned, in terms of both what it represents, and then how to represent it in a manner that will ensure that unethical lethal behavior is not produced. The primary question is how to operationalize information regarding the application of lethality that is available in the LOW and ROE, which prescribes the "what is permissible", and then to determine how to implement it within a hybrid deliberative/reactive robotic architecture. Reiterating from the last section: the set of ethical constraints C defines the space where a lethal action constitutes a valid permissible or obligated response. The application of lethal force as a response <u>must</u> be constrained by the LOW and ROE before it can be employed by the autonomous system.

We are specifically dealing here with “bounded morality” [Allen et al. 06], a system that can adhere to its moral standards within the situations that it has been designed for, in this case specific battlefield missions. It is thus equally important to be able to represent these situations correctly to ensure that the system will indeed provide the appropriate response when encountered. This is further complicated by the variety of sensor and information feeds that are available to a particular robotic implementation. Thus it is imperative that the robot be able to assess the situation correctly in order to respond ethically. A lethal response for an incorrectly identified situation is unacceptable. Clearly this is a non-trivial task. For the majority of this article, however, we will assume that effective situational assessment methods exist, and then given a particular battlefield situation, we examine how an appropriate response can be generated.

This requires determining at least two things: specifically what content we need to represent to ensure the ethical application of lethality (Section 4.1) and then how to represent it (Section 4.2). Section 5 addresses the issues regarding how to put this ethical knowledge to work from an architectural perspective once it has been embedded in the system. Clearly the representational choices that are made will significantly affect the overall architectural design.

4.1 Specific issues for lethality – What to represent

The application of lethality by a robot in one sense is no different than the generation of any particular robotic response to a given situation. In our view, however, we chose to designate the actions with potential for lethality as a class of special **privileged** responses which are governed by a set of external factors, in this case the Laws of War and other related ethical doctrine such as the Rules of Engagement.

Issues surround the underpinning ethical structure, i.e., whether a utilitarian approach is applied, which can afford a specific calculus for the determination of action (e.g., [Brandt 82, Cloos 05]), or a deontological basis that invokes a rights or duty-based approach (e.g., [Powers 05]). This will impact the selection of the representations to be chosen. Several options are described below in support of the decision regarding the representations to be employed in the architecture outlined in Section 5.

While robotic responses in general can be encoded using either discrete or continuous approaches as mentioned in Section 3, for behaviors charged with the application of weapons they will be considered as a binary discrete response ($\mathbf{r}$), i.e., the weapon system is either fired with intent or not. There may be variability in a range of targeting parameters, some of which involve direct lethal intent and others that do not, such as weapon firing for warning purposes (a shot across the bow), probing by fire (testing to see if a target is armed or not), reconnaissance by fire (searching for responsive combatant targets using weaponry), wounding with non-lethal intent, or deliberate lethal intent. There may also be variations in the patterns of firing both spatially and temporally (e.g., single shot, multiple bursts with pattern, suppressing fire, etc.) but each of these will be considered as separate discrete behavioral responses $\mathbf{r_{ij}}$, all of which, nonetheless, have the potential effect of resulting in lethality, even if unintended. The application of non-lethal weaponry, e.g., tasers, sting-nets, foaming agents etc., also can be considered as discrete responses, which although are technically designated as non-lethal responses can also potentially lead to unintentional lethality.

4.1.1 Laws of War

But specifically what are we trying to represent within the architecture? Some examples can be drawn from the United States Army Field Manual FM 27-10 *The Law of Land Warfare* [US Army 56], which states that the law of land warfare:

> *"is inspired by the desire to diminish the evils of war by*
>
> *a) protecting both combatants and noncombatants from unnecessary suffering;*
>
> *b) safeguarding certain fundamental human rights of persons who fall into the hands of the enemy, particularly prisoners of war, the wounded and sick, and civilians; and*
>
> c) *Facilitating the restoration of peace."*

Although lofty words, they provide little guidance regarding specific constraints. Other literature can help us in that regard. [Waltzer 77, pp 41-42] recognizes two general classes of prohibitions that govern the "central principle that soldiers have an equal right to kill. … War is distinguishable from murder and massacre only when restrictions are established on the reach of the battle". The resulting restrictions constitute the set of constraints *C* we desire to represent.

The underlying principles that guide modern military conflict are [Bill 00]:

1. **Military Necessity**: One may target those things which are not prohibited by LOW and whose targeting will produce a military advantage. Military Objective: persons, places, or objects that make an effective contribution to military action.
2. **Humanity or Unnecessary Suffering**: One must minimize unnecessary suffering incidental injury to people and collateral damage to property.
3. **Proportionality**: The US Army prescribes the test of proportionality in a clearly utilitarian perspective as: "The loss of life and damage to property incidental to attacks must not be <u>excessive</u> in relation to the <u>concrete and direct</u> military advantage expected to be gained." [US Army 56 , para. 41, change 1]
4. **Discrimination or Distinction**: One must discriminate or distinguish between combatants and non-combatants, military objectives and protected people/protected places.

These restrictions determine *when and how* soldiers can kill and *who* they can kill. Specific U.S. Army policy assertions from Army headquarters Field Manual FM3-24 validate the concepts of lawful warfighting [USArmy 06]:

- Combat, including COIN [Counterinsurgency] and other irregular warfare, often obligates Soldiers and Marines to choose the riskier course of action to minimize harm to noncombatants.
- Even in conventional operations, Soldiers and Marines are not permitted to use force disproportionately or indiscriminately.
- As long as their use of force is proportional to the gain to be achieved and discriminate in distinguishing between combatants and noncombatants, Soldiers and Marines may take actions where they knowingly risk, but do not intend, harm to noncombatants. [Principle of Double Effect]

- Combatants must take all feasible precautions in the choice of means and methods of attack to avoid and minimize loss of civilian life, injury to civilians, and damage to civilian objects.

Drawing directly from the Laws of War, we now aggregate specific prohibitions, permissions, and obligations that the warfighter (and an ethical autonomous system) must abide by. It must be ensured that these constraints are effectively embedded within a robot potentially capable of lethal action for the specific battlefield situations it will encounter.

Specific examples of prohibited acts include [US Army 56]:

1. It is especially forbidden
    a. To declare that no quarter will be given the enemy.
    b. To kill or wound an enemy who, having laid down his arms, or having no longer means of defense, has surrendered at discretion.
    c. To employ arms, projectile, or material calculated to cause unnecessary suffering.
2. The pillage of a town or place, even when taken by assault is prohibited.
3. The taking of hostages is prohibited (including civilians).
4. Devastation as an end in itself or as a separate measure of war is not sanctioned by the law of war. There must be some reasonably close connection between the destruction of property and the overcoming of the enemy's army.

Regarding lawful targeting (who can and cannot be killed and what can be targeted in warfare):

1. *Regarding combatants and military objectives*:
    a. Once war has begun, soldiers (combatants) are subject to attack at any time, unless they are wounded or captured. [Waltz 77, p. 138]
    b. Targeting of enemy personnel and property is permitted unless otherwise prohibited by international law. [Bill 00, p. 152]
    c. Attacks on military objectives which may cause collateral damage to civilian objects or collateral injury to civilians not taking a direct part in the hostilities are not prohibited (Principle of Double Effect). [Rawcliffe and Smith 06, p. 21]
    d. Collateral/Incidental damage is not a violation of international law in itself (subject to the law of proportionality). [Bill 00, p. 154]
    e. All reasonable precautions must be taken to ensure only military objectives are targeted, so damage to civilian objects (collateral damage) or death and injury to civilians (incidental injury) is avoided as much as possible. [Klein 03]
    f. The presence of civilians in a military objective does not alter its status as a military objective. [Rawcliffe and Smith 06, p. 23]
    g. In general, any place the enemy chooses to defend makes it subject to attack. This includes forts or fortifications, places occupied by a combatant force or through

which they are passing, and city or town with indivisible defensive positions. [Bill 00 p. 160]

h. A belligerent attains combatant status by merely carrying his arms openly during each military engagement, and visible to an adversary while deploying for an attack. (The United States believes this is not an adequate test as it "diminishes the distinction between combatants and civilians, thus undercutting the effectiveness of humanitarian law"). [Bill 00, 157]

i. Retreating troops, even in disarray, are legitimate targets. They could only be immunized from further attack by surrender, not retreat. [Dinstein 02]

j. Destroy, take or damage property based <u>only</u> upon military necessity. [Bill 00, 140]

k. A fighter must wear "a fixed distinctive sign visible at a distance" and "carry arms openly" to be eligible for the war rights of soldiers. Civilian clothes should not be used as a ruse or disguise. [Waltzer 77, p. 182]

l. [Dinstein 02] enumerates what he views as legitimate military objectives under the current *Jus in Bello*:

   1) Fixed military fortifications, bases, barracks and installations, including training and war-gaming facilities
   2) Temporary military camps, entrenchments, staging areas, deployment positions, and embarkation points
   3) Military units and individual members of the armed forces, whether stationed or mobile
   4) Weapon systems, military equipment and ordnance, armor and artillery, and military vehicles of all types
   5) Military aircraft and missiles of all types
   6) Military airfields and missile launching sites
   7) Warships (whether surface vessels or submarines) of all types
   8) Military ports and docks
   9) Military depots, munitions dumps, warehouses or stockrooms for the storage of weapons, ordnance, military equipment and supplies (including raw materials for military use, such as petroleum)
   10) Factories (even when privately owned) engaged in the manufacture of arms, munitions and military supplies
   11) Laboratories or other facilities for the research and development of new weapons and military devices
   12) Military repair facilities
   13) Power plants (electric, hydroelectric, etc.) serving the military

14) Arteries of transportation of strategic importance, principally mainline railroads and rail marshaling yards, major motorways, navigable rivers and canals (including the tunnels and bridges of railways and trunk roads)

15) Ministries of Defense and any national, regional or local operational or coordination center of command, control and communication relating to running the war (including computer centers, as well as telephone and telegraph exchanges, for military use)

16) Intelligence-gathering centers (even when not run by the military establishment)

17) All enemy warships

18) An enemy merchant vessel engaged directly in belligerent acts (e.g., laying mines or minesweeping)

19) An enemy merchant vessel acting as an auxiliary to the enemy armed forces (e.g., carrying troops or replenishing warships)

20) An enemy merchant vessel engaging in reconnaissance or otherwise assisting in intelligence gathering for the enemy armed forces

21) An enemy merchant vessel refusing an order to stop or actively resisting capture

22) An enemy merchant vessel armed to an extent that it can inflict damage on a warship (especially a submarine)

23) An enemy merchant vessel traveling under a convoy escorted by warships, thereby benefiting from the (more powerful) armament of the latter

24) An enemy merchant vessel making an effective contribution to military action (e.g., by carrying military materials)

25) All enemy military aircraft

26) Enemy civilian aircraft when flying within the jurisdiction of their own State, should enemy military aircraft approach and they do not make the nearest available landing

27) Enemy civilian aircraft when flying (i) within the jurisdiction of the enemy; or (ii) in the immediate vicinity thereof and outside the jurisdiction of their own State; or (iii) in the immediate vicinity of the military operations of the enemy by land or sea (the exceptional right of prompt landing is inapplicable)

2. *Regarding noncombatant immunity:*

a. Civilians**:**

1) Individual civilians, the civilian population as such and civilian objects are protected from intentional attack. [Rawcliffe and Smith 06, p. 23]

2) Civilians are protected from being sole or intentional objects of a military attack, from an indiscriminate attack, or attack without warning prior to a bombardment [Bill 00, p. 157] unless and for such time as he or she takes a direct part in hostilities. [Rawcliffe and Smith 06, p. 29]

3) Launching attacks against civilian populations is prohibited [Klein 03]. Noncombatants cannot be attacked at any time or be the targets of military activity (noncombatant immunity). [Waltz 77 p. 153]

4) There exists an obligation to take feasible measures to remove civilians from areas containing military objectives. [Bill 00, p. 136]

5) It is forbidden to force civilians to give information about the enemy. [Brandt 72]

6) It is forbidden to conduct reprisals against the civilian population "on account of the acts of individuals for which they cannot be regarded as jointly and severally responsible". [Brandt 72]

7) Treatment of C¥civilians [Bill 00, pp. 129-130,139-141] (including those in conflict area):

    a) No adverse distinction based upon race, religion, sex, etc.

    b) No violence to life or person

    c) No degrading treatment

    d) No civilian may be the object of a reprisal

    e) No measures of brutality

    f) No coercion (physical or moral) to obtain information

    g) No insults and exposure to public curiosity

    h) No general punishment for the acts of an individual, subgroup, or group

    i) Civilians may not be used as "human shields" in an attempt to immunize an otherwise lawful military objective. However, violations of this rule by the party to the conflict do not relieve the opponent of the obligation to do everything feasible to implement the concept of distinction (discrimination)

    j) Civilian wounded and sick must be cared for

    k) Special need civilians are defined as: mothers of children under seven; wounded, sick and infirm; aged; children under the age of 15; and expectant Mothers; which results from the presumption that they can play no role in support of the war effort. Special need civilians are to be respected and protected by all parties to the conflict at all times. This immunity is further extended to Ministers, medical personnel and transport, and civilian hospitals.

8) In order to ensure respect and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly direct their operations only against military objectives [UN 48]. This includes the following specific prohibitions:

    a) Civilians may never be the object of attack.

    b) Attacks intended to terrorize the civilian population are prohibited.

        c) Indiscriminate attacks are prohibited. Indiscriminate is defined as:

        (1) Attacks not directed at a specific military objective, or employing a method or means of combat that cannot be so directed

        (2) Attacks which employ a method or means of combat the effects of which cannot be controlled

        (3) Attacks treating dispersed military objectives, located in a concentration of civilians, as one objective

        (4) Attacks which may be expected to cause collateral damage excessive in relation to the concrete and direct military advantage to be gained (proportionality)

b. Prisoners of War (POWs) [Bill 00 p, 158]:

    1) Surrender may be made by any means that communicates the intent to give up (no clear rule)
    2) Onus is on person or force surrendering to communicate intent to surrender
    3) Captor must not attack, and must protect those who surrender (no reprisals)
    4) A commander may not put his prisoners to death because their presence retards his movements or diminishes his power of resistance by necessitating a large guard…or it appears that they will regain their liberty through the impending success of their forces. It is likewise unlawful for a commander to kill his prisoners on the grounds of self-preservation. [US Army 56]

c. Medical personnel, relief societies, religious personnel, journalists, and people engaged in the protection of cultural property shall not be attacked. [Bill 00 p. 159]

d. Passing sentences and carrying out [summary] executions without previous judgment of a regularly constituted court is prohibited at any time and in any place whatsoever. [US Army 04, p. 3-38]

3. *<u>Regarding non-military objectives</u>:*

    a. A presumption of civilian property attaches to objects traditionally associated with civilian use (dwellings, schools, etc.) as contrasted with military objectives, i.e., they are presumed not subject to attack. [Rawcliffe and Smith 06, p. 23]

    b. Undefended places are not subject to attack. This requires that all combatants and mobile military equipment be removed, no hostile use of fixed military installations, no acts of hostility committed by the authorities or the population, no activities in support of military operations present (excluding medical treatment and enemy police forces). [Bill 00 160]

    c. The environment cannot be the object of reprisals. Care must be taken to prevent long-term, widespread and severe damage. [Bill 00, 161]

    d. Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science, charitable purposes, and historic monuments. The enemy has a duty to mark them clearly with visible and distinctive signs. Misuse will make them subject to attack. [Bill 00 p. 162]

e. Works and installations containing dangerous forces should be considered to be immune from attack. This includes nuclear power plants, dams, dikes, etc. (This is not U.S. law, however, which believes standard proportionality test should apply).

f. It is prohibited to attack, destroy, remove or render useless objects indispensable for survival of the civilian population, such as foodstuffs, crops, livestock, water installations, and irrigation works [Rawcliffe and Smith 06, p. 24] unless these objects are used <u>solely</u> to support the enemy military. [Bill 00, p. 135]

g. There exists an obligation to take feasible precautions in order to minimize harm to non-military objectives. [Bill 00, p. 135]

4. *<u>Regarding Use of Arms:</u>*

a. Cannot use lawful arms in a manner that causes unnecessary suffering or used with the intent to cause civilian suffering (proportionality). The test essentially is whether the suffering occasioned by the use of the weapon is needless, superfluous, or grossly disproportionate to the advantage gained by its use. [Bill 00, p. 164]

b. Indiscriminate attacks are prohibited. This includes attacks not directed against a military objective and the use of a method of attack that cannot be effectively directed or limited against an enemy objective. [Bill 00, p. 154-156]

5 *<u>Regarding War Crime Violations:</u>*

a. All violations of the law of war should be promptly reported to a superior. [US Army 06, Rawcliffe and Smith 06, p. 45]

b. Members of the armed forces are bound to obey only lawful orders. [US Army 04, p. 3-37]

c. Soldiers must also attempt to prevent LOW violations by other U.S. soldiers. [Rawcliffe and Smith 06, p. 45]

d. (Troop Information) In the rare case when an order seems unlawful, don't carry it out right away but don't ignore it either, instead seek clarification of that order. [Rawcliffe and Smith 06, p. 43]

6. *<u>Regarding definition of civilians</u>*:

An important issue regarding discrimination is how to determine who is defined as a civilian [Bill 00, pp. 127-129] to afford them due protection from war. As late as 1949, the fourth Geneva Convention, which was primarily concerned with the protection of civilians, provided no such definition and relied on common sense, which may be hard to operationalize in modern warfare. In the 1977 Protocol I commentary, it was acknowledged that a clear definition is essential, but used an awkward negative definition: anyone who does not qualify for Prisoner of War (POW) status, i.e., does not have combatant status, is considered a civilian. This is clarified further by the following [US Army62]:

> *The immunity afforded individual civilians is subject to an overriding condition, namely, on their abstaining from all hostile acts. Hostile acts should be understood to be acts which by their nature and purpose are intended to cause actual harm to the personnel and equipment of the armed forces.* ***Thus a civilian who takes part in armed combat, either individually or as part of a group, thereby becomes a legitimate target*** *. . .*

Expanding further: "This 'actual harm' standard is consistent with contemporary U.S. practice, as reflected in ROE-based 'harmful act/harmful intent' test for justifying use of deadly force against civilians during military operations." [Bill 00]

Those civilians who participate only in a general sense in the war effort (non-hostile support, manufacturing, etc.) are excluded from attack [Bill 00, US Army 56]: "According to Article 51(3) [Geneva Convention Protocol I of 1977], civilians shall enjoy the protection of this section (providing general protection against dangers arising from military operations) **unless and for such time as they take a direct part in hostilities**." … where "direct part" means acts of war which by their nature or purpose are likely to cause actual harm to the personnel and equipment of the enemy armed forces. .... Although the United States decided not to ratify Protocol I, there was no indication that this definition of "civilian" was objectionable.

Appendix A contains the specific language used in the U.S. Military manual that describes these Laws of War in more detail. We will restrict ourselves in this research to those laws that are specifically concerned with the application of lethality in direct combat, but it is clear that a more expansive treatment of ethical behavior of autonomous systems should also be considered in the future.

4.1.2 Rules of Engagement

In order to provide more mission and context-sensitive guidance regarding the use of force in the battlefield, Rules of Engagement (ROE), Rules for the Use of Force (RUF) and General Orders are provided in advance of an engagement [US Army 04]. "United States soldiers and marines face hard choices about what, when, and where they can shoot" [Martins 94]. Rules of engagement are concerned with when and where military force may be used and against whom and how it should be used. ROE are drafted in conjunction with Judge Advocates with the intent that they are legally and tactically sound, versatile, understandable and easily executed [Berger et al. 04]. Rules of engagement are defined as follows:

> *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.* [DOD 07]

Two high-level functions of the ROE are to provide guidance from the President and Secretary of Defense to *deployed units* on the use of force and to act as a control mechanism for the transition from peacetime to combat operations (war) [Berger et al. 04]. Ten specific ROE function types include (from [Martins 94]):

1. **Hostility Criteria** - Provide those making decisions whether to fire with a set of objective factors to assist in determining whether a potential assailant exhibits hostile intent and thus clarify whether shots can be fired before receiving fire.
2. **Scale of Force or Challenge Procedure:** Specify a graduated show of force that ground troops must use in ambiguous situations before resorting to deadly force. Include such measures as giving a verbal warning, using a riot stick, perhaps firing a warning shot, or firing a shot intended to wound. May place limits on the pursuit of an attacker.
3. **Protection of Property and Foreign Nationals:** Detail what and who may be defended with force aside from the lives of United States soldiers and citizens. May include measures to be taken to prevent crimes in progress or the fleeing of criminals. May place limits on pursuit of an attacker.
4. **Weapon Control Status or Alert Conditions:** Announce, for air defense assets, a posture for resolving doubts over whether to engage. Announce, for units observing alert conditions, a series of measures designed to adjust unit readiness for attack to the level of the perceived threat. The measures may include some or all of the other functional types of rules.
5. **Arming Orders:** Dictate which soldiers in the force are armed and which have live ammunition. Specify which precise orders given by whom will permit the loading and charging of firearms.
6. **Approval to Use Weapons Systems:** Designates what level commander must approve use of particular weapons systems. Perhaps prohibits use of a weapon entirely.
7. **Eyes on Target:** Require that the object of fire be observed by one or more human or electronic means.

8. **Territorial or Geographic Constraints:** Create geographic zones or areas into which forces may not fire. May designate a territorial, perhaps political boundary, beyond which forces may neither fire nor enter except perhaps in hot pursuit of an attacking force. Include tactical control measures that coordinate fire and maneuver by means of graphic illustrations on operations map overlays, such as coordinated fire lines, axes of advance, and direction of attack.
9. **Restrictions on Manpower:** Prescribe numbers and types of soldiers to be committed to a theatre or area of operations. Perhaps prohibit use of United States manpower in politically or diplomatically sensitive personnel assignments requiring allied manning.
10. **Restrictions on Point Targets and Means of Warfare**: Prohibit targeting of certain individuals or facilities. May restate basic rules of the Law of War for situations in which a hostile force is identified and prolonged armed conflict ensues.

Standing Rules of Engagement

There are both Standing Rules of Engagement (SROE), which are global in context, applying to all missions, and ROE which are customized for the needs of the mission. All are intended to strictly adhere to the LOW. The following definitions are used for the SROE [Berger et al. 04]:

a) **Hostile Act:** An attack or other use of force against the United States, U.S. forces, and, in certain circumstances, U.S. nationals, their property, U.S. commercial assets, and/or other designated non-U.S. forces, foreign nationals and their property. It is also force used directly to preclude or impede the mission and/or duties of U.S. forces, including the recovery of U.S. personnel and vital U.S. Government property. A hostile act triggers the right to use *proportional force* in self defense to deter, neutralize or destroy the threat.

b) **Hostile Intent:** The threat of imminent use of force against the United States, U.S. forces, or other designated persons and property. It is also the threat of force used directly to preclude or impede the mission and/or duties of U.S. forces, including the recovery of U.S. personnel and vital U.S. Government property. When hostile intent is present, the right exists to use *proportional force* in self defense to deter, neutralize or destroy the threat.

c) **Hostile Force:** Any civilian, paramilitary, or military force or terrorist(s), with or without national designation, that has committed a hostile act, exhibited hostile intent, or has been declared hostile by appropriate U.S. authority.

d) **Declaring Forces Hostile:** Once a force is declared to be "hostile," U.S. units may engage it without observing a hostile act or demonstration of hostile intent; i.e., the basis for engagement shifts from conduct to status. The authority to declare a force hostile is limited.

e) **Necessity**: when a hostile act occurs or when a force or terrorists exhibits hostile intent.

f) **Proportionality**: Force used to counter a hostile act or demonstrated hostile intent must be reasonable in intensity, duration, and magnitude to the perceived or demonstrated threat based on all facts known to the commander at the time.

SROE focus on self-defense, i.e., “a commander may use the weapon of choice, unless specifically prohibited, tempered only by proportionality and necessity” [Womack 96]. Self-defense is considered in the context of the nation, collective (Non-US entities), unit, and of the individual.

The SROE permissible actions for self-defense are stated clearly [Berger el al 04] and the relevant ones are reproduced below:

Means of Self-Defense. All necessary means available and all appropriate actions may be used in self-defense. The following guidelines apply for individual, unit, national, or collective self-defense:

> (1) Attempt to De-Escalate the Situation. When time and circumstances permit, the hostile force should be warned and given the opportunity to withdraw, or cease threatening actions.
>
> (2) Use Proportional Force -- Which May Include Nonlethal Weapons – to Control the Situation. When the use of force in self-defense is necessary, the nature, duration, and scope of the engagement should not exceed that which is required to decisively counter the hostile act or demonstrated hostile intent and to ensure the continued protection of U.S. forces or other protected personnel or property.
>
> (3) Attack to Disable or Destroy. An attack to disable or destroy a hostile force is authorized when such action is the only prudent means by which a hostile act or demonstration of hostile intent can be prevented or terminated. When such conditions exist, engagement is authorized only while the hostile force continues to commit hostile acts or exhibit hostile intent.

Pursuit of Hostile Forces. Self-defense includes the authority to pursue and engage hostile forces that continue to commit hostile acts or exhibit hostile intent.

The Standing Rules of Engagement (SROE) provide for implementation and guidance on the right and obligation of self-defense and the application of force for mission accomplishment. “The SROE do not limit a commander’s inherent authority and obligation to use all necessary means available to take all appropriate action in self-defense of the commander’s unit and other U.S. forces in the vicinity” [AFJAGS 06]. Hot pursuit in self-defense is permissible, where an enemy force can be pursued and engaged that has either committed a hostile act or demonstrated hostile intent and remains an imminent threat [SROE 94].

<u>Rules of Engagement (non-SROE)</u>

Supplemental ROE measures are applicable beyond the SROE.

> “The current SROE now recognizes a fundamental difference between the supplemental measures. Those measures that are reserved to the President or Secretary of Defense or Combatant Commander are generally **restrictive**, that is, either the President or Secretary of Defense or Combatant Commander must specifically permit the particular operation, tactic, or weapon before a field commander may utilize them. Contrast this with the remainder of the supplemental measures, those delegated to subordinate commanders. These measures are all **permissive** in nature, *allowing a commander to use any weapon*

> *or tactic available and to employ reasonable force to accomplish his mission*, without having to get permission first. Inclusion within the subordinate commanders' supplemental list does not suggest that a commander needs to seek authority to use any of the listed items. *SUPPLEMENTAL ROE RELATE TO MISSION ACCOMPLISHMENT, NOT TO SELF-DEFENSE, AND NEVER LIMIT A COMMANDER'S INHERENT RIGHT AND OBLIGATION OF SELF DEFENSE".* [Berger et al. 04]

We can use this notion of restrictive and permissive measures (we will use the stronger version of *obligated* instead of *permissive*) to advantage in the design of representations and architectural methods to be developed for use in lethal autonomous systems as described in subsequent sections of this article.

Every operations plan normally provides ROE as part of the mission. They are different for each operation, area, and can change as the situation changes. There are classified ROE documents that provide general guidance for specific air, land, and sea mission operations. There also exist Theater-Specific ROE for use by Combatant Commanders in the Area of Responsibility that address strategic and political sensitivities.

ROE are tailored to local circumstances, the nature and history of the threat, and must be dynamic and changing as the mission evolves [AFJAGS 06]. They do not limit a soldier's right to self-defense. "The ROE are frequently more restrictive than the Law of War, because they take into consideration the specifics of the operating environment, such as culture, religious sensitivities, geography, historical monuments, and so forth" [USM 07]. They are based upon LOW, US foreign policy, US domestic law and concerns, and operational matters. Military necessity for self-defense requires that a hostile act occur or the exhibition of hostile intent before armed force is permitted. Proportionality states that the force used must have intensity, duration, and magnitude that is reasonable based upon the information available at the time. No more force than is necessary is to be employed.

[Sagan 91] observes 2 types of ROE failures that can occur in their writing. A ROE *Weakness* error occurs when the rules are excessively tight, so a commander cannot effectively complete his mission or defeat an attack. A ROE *Escalatory* error occurs if the rules are excessively loose to the point where force may be used that is deemed undesirable by political authorities (it should never be illegal). Great care should be taken in the writing of the ROE for lethal autonomous systems to avoid both failure types, but especially escalatory ones.

ROE can be in the form of a *command by negation* where a soldier can act on his own in this manner unless explicitly forbidden, or a positive command which can only be taken if explicitly ordered by a superior. ROE are deliberated upon well in advance of an engagement, may cover several scenarios and have different rules for each [Wikipedia 07b].

Some of the basics of the ROE include (forming the RAMP acronym) [US Army 04]:

- <u>Return Fire with Aimed Fire.</u> Return force with force. You always have the right to repel hostile acts with necessary force.
- <u>Anticipate</u> Attack. Use force if, but only if, you see clear indicators of hostile intent.
- <u>Measure</u> the amount of force that you use, if time and circumstances permit. Use only the amount of force necessary to protect lives and accomplish the mission.
- <u>Protect</u> with deadly force only human life, and property designated by your commander. Stop short of deadly force when protecting other property.

Several example ROE Cards appear on the following pages (from [Berger et al. 04]). A comprehensive list of ROE cards and vignettes is available in [CLAMO 00].

An Example for **Armed Conflict** (War) follows:

DESERT STORM
RULES OF ENGAGEMENT

ALL ENEMY MILITARY PERSONNEL AND VEHICLES TRANSPORTING THE ENEMY OR THEIR SUPPLIES MAY BE ENGAGED SUBJECT TO THE FOLLOWING RESTRICTIONS:

A. Do not engage anyone who has surrendered, is out of battle due to sickness or wounds, is shipwrecked, or is an aircrew member descending by parachute from a disabled aircraft.

B. Avoid harming civilians unless necessary to save U.S. lives. Do not fire into civilian populated areas or buildings which are not defended or being used for military purposes.

C. Hospitals, churches, shrines, schools, museums, national monuments, and other historical or cultural sites will not be engaged except in self defense.

D. Hospitals will be given special protection. Do not engage hospitals unless the enemy uses the hospital to commit acts harmful to U.S. forces, and then only after giving a warning and allowing a reasonable time to expire before engaging, if the tactical situation permits.

E. Booby traps may be used to protect friendly positions or to impede the progress of enemy forces. They may not be used on civilian personal property. They will be recovered and destroyed when the military necessity for their use no longer exists.

F. Looting and the taking of war trophies are prohibited.

G. Avoid harming civilian property unless necessary to save U.S. lives. Do not attack traditional civilian objects, such as houses, unless they are being used by the enemy for military purposes and neutralization assists in mission accomplishment.

H. Treat all civilians and their property with respect and dignity. Before using privately owned property, check to see if publicly owned property can substitute. No requisitioning of civilian property, including vehicles, without permission of a company level commander and without giving a receipt. If an ordering officer can contract the property, then do not requisition it.

I. Treat all prisoners humanely and with respect and dignity.

J. ROE Annex to the OPLAN provides more detail. Conflicts between this card and the OPLAN should be resolved in favor of the OPLAN.

REMEMBER

1. FIGHT ONLY COMBATANTS.
2. ATTACK ONLY MILITARY TARGETS.
3. SPARE CIVILIAN PERSONS AND OBJECTS.
4. RESTRICT DESTRUCTION TO WHAT YOUR MISSION REQUIRES.

ROE Card for a marine **operation other than war** (OOTW) in an urban environment, in this case for the evacuation of UN Peacekeeping troops in Somalia in 1995:

ROE Used for Operation United Shield

Nothing in these Rules of Engagement limits your right to take appropriate action to defend yourself and your unit.

a. You have the right to use deadly force in response to a hostile act or when there is a clear indication of hostile intent.

b. Hostile fire may be returned effectively and promptly to stop a hostile act.

c. When US forces are attacked by unarmed hostile elements, mobs and/or rioters, US forces should use the minimum force necessary under the circumstances and proportional to the threat.

d. Inside designated security zones, once a hostile act or hostile intent is demonstrated, you have the right to use minimum force to prevent armed individuals/crew-served weapons from endangering US/UNOSOM II forces. This includes deadly force.

e. Detention of civilians is authorized for security reasons or in self-defense.

Remember:

1. The United States is not at war.
2. Treat all persons with dignity and respect.
3. Use minimum force to carry out mission.
4. Always be prepared to act in self-defense.

An ROE Card example for a **Peacekeeping** mission in Kosovo.

KFOR RULES OF ENGAGEMENT FOR USE IN KOSOVO

MISSION. Your mission is to assist in the implementation of and to help ensure compliance with a Military Technical Agreement (MTA) in Kosovo.

SELF-DEFENSE.

a. You have the right to use necessary and proportional force in self-defense.

b. Use only the minimum force necessary to defend yourself.

GENERAL RULES.

a. Use the minimum force necessary to accomplish your mission.

b. Hostile forces/belligerents who want to surrender will not be harmed. Disarm them and turn them over to your superiors.

c. Treat everyone, including civilians and detained hostile forces/belligerents, humanely.

d. Collect and care for the wounded, whether friend or foe.

e. Respect private property. Do not steal. Do not take "war trophies".

f. Prevent and report all suspected violations of the Law of Armed Conflict to superiors.

CHALLENGING AND WARNING SHOTS.

a. If the situation permits, issue a challenge:

- In **English**: "NATO! STOP OR I WILL FIRE!"
- Or in **Serbo-Croat**: "NATO! STANI ILI PUCAM!"
- (Pronounced as: "NATO! STANI ILI PUTSAM!)
- Or in **Albanian**: "NATO! NDAL OSE UNE DO TE QELLOJ!
- (Pronounced as: "NATO! N'DAL OSE UNE DO TE CHILLOY!)

b. If the person fails to halt, you may be authorized by the on-scene commander or by standing orders to fire a warning shot.

OPENING FIRE.

a. You may open fire only if you, friendly forces or persons or property under your protection are threatened with deadly force. This means:

(1) You may open fire against an individual who fires or aims his weapon at, or otherwise demonstrates an intent to imminently attack, you, friendly forces, or Persons with Designated Special Status (PDSS) or property with designated special status under your protection.

(2) You may open fire against an individual who plants, throws, or prepares to throw, an explosive or incendiary device at, or otherwise demonstrates an intent to imminently attack you, friendly forces, PDSS or property with designated special status under your protection.

(3) You may open fire against an individual deliberately driving a vehicle at you, friendly forces, or PDSS or property with designated special status.

b. You may also fire against an individual who attempts to take possession of friendly force weapons, ammunition, or property with designated special status, and there is no way of avoiding this.

c. You may use minimum force, including opening fire, against an individual who unlawfully commits or is about to commit an act which endangers life, in circumstances where there is no other way to prevent the act.

MINIMUM FORCE.

a. If you have to open fire, you must:

- Fire only aimed shots; and
- Fire no more rounds than necessary; and
- Take all reasonable efforts not to unnecessarily destroy property; and
- Stop firing as soon as the situation permits.

b. You may not intentionally attack civilians, or property that is exclusively civilian or religious in character, except if the property is being used for military purposes or engagement is authorized by the commander.

Rules for the Use of Force

The Rules for the Use of Force (RUF) provides rules for performing security duty **within the United States.** They are escalating rules used as a last resort, and provide for the use of lethal force in the following conditions [US Army 04]:

- For immediate threat of death or serious bodily injury to self or others
- For defense of persons under protection
- To prevent theft, damage, or destruction of firearms, ammunition, explosives, or property designated vital to national security

The escalation should adhere to the following pattern for security [US Army 04]:

1. SHOUT - verbal warning to halt.
2. SHOVE – non-lethal physical force.
3. SHOW - intent to use weapon.
4. SHOOT - deliberately aimed shots until threat no longer exists.

- Warning shots are not permitted.

ROE for Peace Enforcement Missions

1. Peace enforcement missions may have varying degrees of expanded ROE and may allow for the use of force to accomplish the mission (i.e., the use of force beyond that of self-defense.)
2. For Chapter VI United Nations Peacekeeping operations, the use of deadly force is justified only under conditions of extreme necessity (typically self-defense) and as a last resort when all lesser means have failed to curtail the use of violence by the parties involved [Rawcliffe and Smith 06, p. 67].

4.2 Representational Choices – How to Represent

[Anderson et al. 04] state that "there is every reason to believe that ethically sensitive machines can be created. There is widespread acknowledgment, however, about the difficulty associated with machine ethics [Moor 06, McLaren 06]. There are several specific problems [McLaren 05]:

1. The laws, codes, or principles (i.e., rules) are almost always provided in a highly conceptual, abstract level.
2. The conditions, premises or clauses are not precise, are subject to interpretation, and may have different meanings in different contexts.
3. The actions or conclusions in the rules are often abstract as well, so even if the rule is known to apply the ethically appropriate action may be difficult to execute due to its vagueness.
4. The abstract rules often conflict with each other in specific situations. If more than one rule applies it is not often clear how to resolve the conflict.

First order predicate logic and other standard logics based on deductive reasoning are not generally applicable as they operate from inference and deduction, not the notion of obligation. Secondly, controversy exists about the correct ethical framework to use in the first place given the multiplicity of philosophies that exist: Utilitarian, Kantian, Social Contract, Virtue Ethics, Cultural Relativism, and so on.

It is my belief that battlefield ethics are more clear-cut and precise than everyday or professional ethics, ameliorating these difficulties somewhat, but not removing them. For this project a commitment to a framework that is consistent with the LOW and ROE must be maintained, strictly adhering to the rights of noncombatants regarding discrimination (deontological), while considering similar principles for the assessment of proportionality based on military necessity (utilitarian). As stated earlier, it is no mean feat to be able to perform situational awareness in a manner to adequately support discrimination. By starting, however, from a "first, do no harm" strategy, battlefield ethics may be feasible to implement, i.e., do not engage a target until obligated to do so consistent with the current situation, and there exists no conflict with the LOW and ROE. If no obligations are present or potential violations of discrimination and proportionality exist, the system cannot fire. By conducting itself in this manner, it is believed that the ethically appropriate use of constrained lethal force can be achieved by an autonomous system.

The ethical autonomy architecture capable of lethal action will use an action-based approach, where ethical theory (as encoded in the LOW and ROE) informs the agent what actions to undertake. Action-based methods have the following attributes [Anderson et al. 06]:

1. Consistency – the avoidance of contradictions in the informing theory
2. Completeness – how to act in any ethical dilemma
3. Practicality – it should be feasible to execute
4. Agreement with expert ethicist intuition

None of these appear out of reach for battlefield applications. The LOW and ROE are designed to be consistent. They should prescribe how to act in each case, and when coupled with a "first, do no harm" as opposed to a "shoot first, ask questions later" strategy (ideally surgically, to further expand upon the medical metaphor of do no harm), the system should act conservatively in the presence of uncertainty (doubt). Bounded morality assures practicality, as it limits the scope of actions available and the situations in which it is permitted to act with lethal force. Agreement with an expert should be feasible assuming they subscribe to the existing International Protocols governing warfare. This expert agreement is also important for the attribution of responsibility and can play a role in the design of the responsibility advisor using methods such as case-based reasoning (Section 4.2.2).

This section reviews the space of potential choices for representing the necessary constraints on lethal action derived from the LOW and ROE for use within the architecture. An overview of these various methods already in use in the nascent field of machine ethics is provided, and they are assessed for utility within the ethical autonomous robot architecture, leading to design commitments for the system as outlined in Section 5.

[Turilli 07] describes a method by which ethical principles can be transformed into an ethically consistent protocol, i.e., a process which produces the same ethical results independent of the actor (computational agents or human individuals - Figure 4). In our case, his original process will need to be transformed somewhat (Fig. 5), but it can still contribute to the correct development of the set of constraints C that are required for the ethical processing within our architecture.

Ethical judgments on action can be seen to take three primary forms: obligatory (the agent is required to conduct the action based on moral grounds), permissible (the action is morally acceptable but not required), and forbidden (the action is morally unacceptable). [Hauser 06, p. 157] outlines the logical relationship between these action classes:

1. If an action is permissible, then it is potentially obligatory but not forbidden.
2. If an action is obligatory, it is permissible and not forbidden.
3. If an action if forbidden, it is neither permissible nor obligatory.

Lethal actions for autonomous systems can potentially fall into any of these classes. Certainly the agent should never conduct a forbidden lethal action, and although an action may be permissible, it should also be deemed obligatory in the context of the mission (military necessity) to determine whether or not it should be undertaken. So in this sense, I argue that any lethal action undertaken by an unmanned system must be obligatory and not solely permissible, where the mission ROE define the situation-specific lethal obligations of the agent and the LOW define absolutely forbidden lethal actions. Although it is conceivable that permissibility alone for the use of lethality is adequate, we will require the provision of additional mission constraints explicitly informing the system regarding target requirements (e.g., as part of the ROE) to define exactly what constitutes an acceptable action in a given mission context. This will also assist with the assignment of responsibility for the use of lethality (Section 5.2.4). Summarizing:

- Laws of War and related ROE determine what are absolutely forbidden lethal actions.
- Rules of Engagement mission requirements determine what is obligatory lethal action, i.e., where and when the agent must exercise lethal force. Permissibility alone is inadequate.

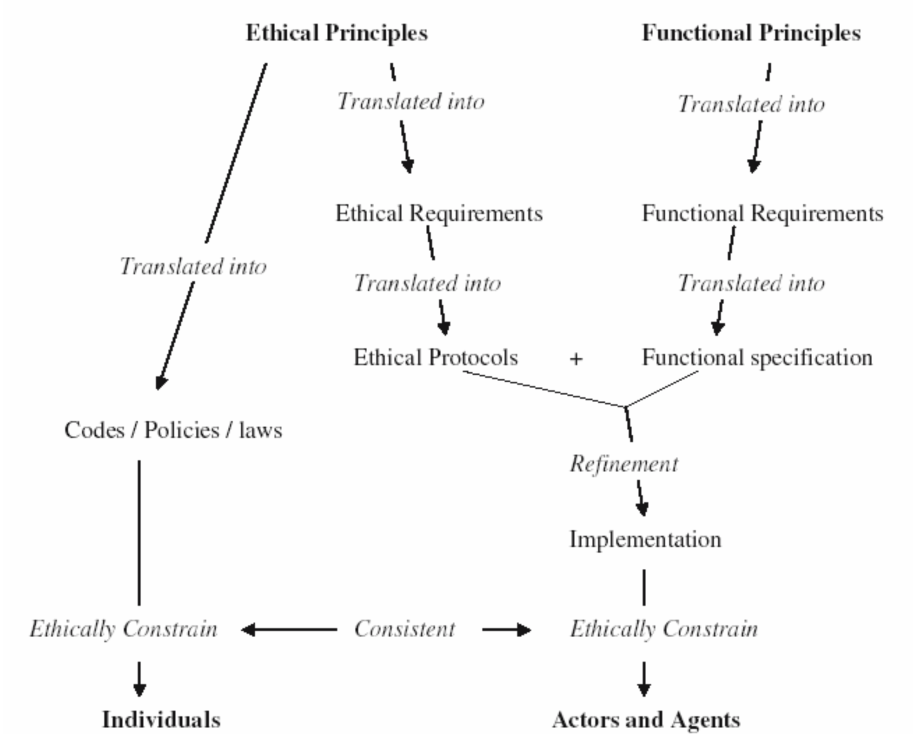


Figure 4: Method for Developing an Ethically Consistent Protocol (from [Turilli 07]).

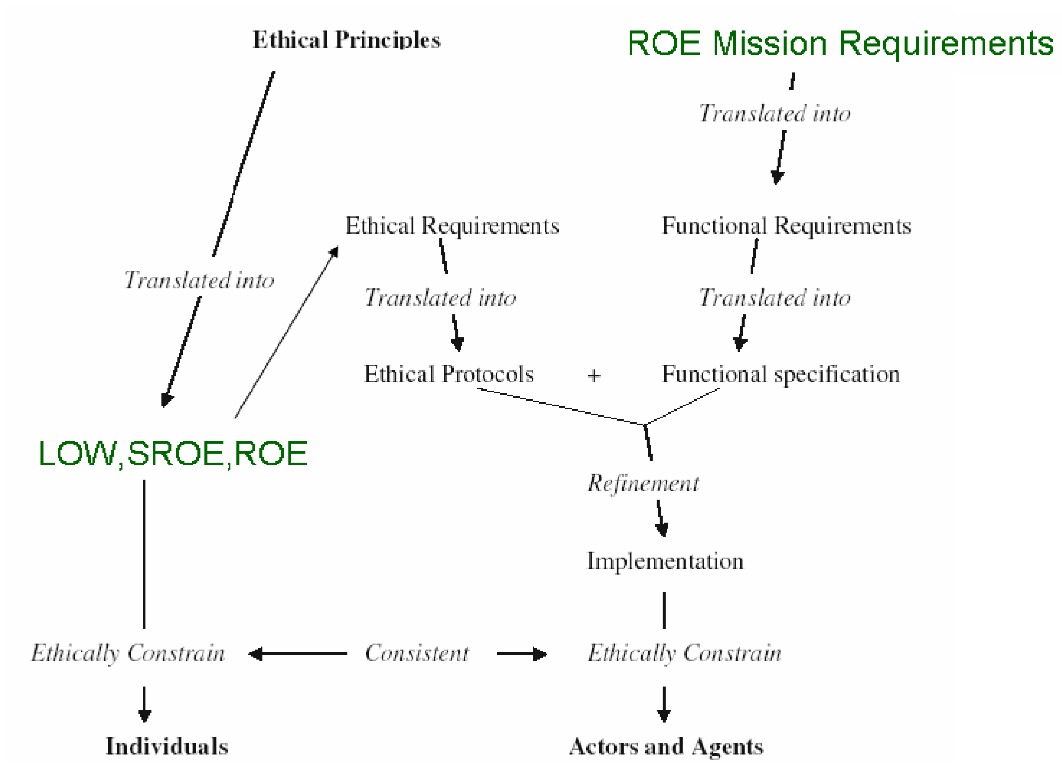


Figure 5: Process for Deriving Constraint set *C* from LOW and ROE.

Let us now relate this back to the set theoretic description in Figures 2-3.

1. Obligatory lethal actions represent $\boldsymbol{P}_{l\text{-}ethical}$ under these restrictions, i.e., the set of ethical lethal actions.
2. Forbidden lethal actions are defined as $\boldsymbol{P}_{l\text{-}unethical} = \boldsymbol{P}_{lethal} - \boldsymbol{P}_{l\text{-}ethical}$, which defines the set of unethical lethal actions.
3. For a lethal response $\boldsymbol{\rho}_{lethal\text{-}ij}$ to be an ethical lethal action $\boldsymbol{\rho}_{l\text{-}ethical\text{-}ij}$ for situation **i**, it must not be forbidden by constraints derived from the LOW, and it must be obligated by constraints derived from the ROE.

It is now our task to:

1. Determine how to represent the LOW as a suitable set of forbidding constraints $C_{Forbidden}$ on $\boldsymbol{P}_{lethal}$ such that any action $\boldsymbol{\rho}_{lethal\text{-}ij}$ produced by the autonomous system is not an element of $\boldsymbol{P}_{l\text{-}unethical}$; and
2. Determine how to represent ROE as a suitable set of obligating constraints $C_{Obligate}$ on $\boldsymbol{P}_{lethal}$ such that any action $\boldsymbol{\rho}_{lethal\text{-}ij}$ produced by the autonomous system is an element of $\boldsymbol{P}_{l\text{-}ethical}$.

Item (1) permits the generation of only non-lethal or ethical lethal (permissible) actions by the autonomous system, and forbids the production of unethical lethal action. Item (2) requires that any lethal action must be obligated by the ROE to be ethical. This aspect of obligation will also assist in the assignment of responsibility, which will be discussed in Section 5.2.4.

Regarding representation for the ethical constraints C, where $C = C_{Forbidden} \cup C_{Obligate}$, there are at least two further requirements:

1. Adequate expressiveness for a computable representation of the ethical doctrine itself.
2. A mechanism by which the representation of the ethical doctrine can be transformed into a form usable within a robotic controller to suitably constrain its actions.

Recalling from Section 3.2, a particular c_k can be considered either:

1. a negative behavioral constraint (a prohibition) that prevents or blocks a behavior $\beta_{lethal\text{-}i}$ from generating $\mathrm{r}_{\text{lethal-ij}}$ for a given perceptual situation S_{j}.
2. a positive behavioral constraint (an obligation) which requires a behavior $\beta_{lethal\text{-}i}$ to produce $\mathrm{r}_{\text{l-ethical-ij}}$ in a given perceptual situational context S_{j}.

It is desirable to have a representation that supports growth of the architecture, where constraints can be added incrementally. This means that we can initially represent a small set of forbidden and obligated constraints and test the overall system without the necessity of a fully complete set of representational constraints that captures the entire space of the LOW and ROE. An underlying assumption will be made that any use of lethality by the autonomous unmanned system is prohibited by default, unless an obligating constraint requires it and it is not in violation of any and all forbidding constraints. This will enable us to incrementally enumerate obligating constraints and be able to assess discrimination capabilities and proportionality evaluation in a step-by-step process. Keep in mind that this project represents only the most preliminary steps towards the design of a fieldable ethical system, and that substantial additional basic and applied research must be conducted before they can even be considered for use in a real world battlefield scenario. But baby steps are better than no steps towards enforcing ethical behavior in autonomous system warfare assuming, as we did in Section 1, its inevitable introduction.

We now review some of the existing approaches that have been applied to the general area of machine ethics and consider their applicability in light of the requirements for representational choices for robotic systems employing lethality consistent with battlefield ethics. It has been observed that there are two major approaches to moral reasoning in the machine ethics community. The first uses moral principles such as exceptionless standards or contributory principles, and is referred to as *generalism*. Exceptionless standards appear to have utility in our context as they [Guarini 06]:

- Specify sufficient conditions for what makes a state of affairs (including actions) good, bad, right, wrong, permissible, impermissible, etc.
- Explain or inform why the principle applies when it does.
- Serve as premises in moral deliberations.

The second approach to moral reasoning is case-based and is referred to as particularism. These different approaches will find compatibility within different places in the autonomous agent architecture capable of lethal action described later: generalism for the run-time reasoning from principles derived from the LOW and ROE, and particularism for the pre-mission role of advising the operator and commanders regarding their responsibility for the use of an agent capable of lethality under a given set of conditions, i.e., a particular case.

4.2.1 *Generalism* – Reasoning from Moral Principles

Most ethical theories, Deontological or Kantian, Utilitarian, Virtue Ethics, etc., assert that an agent should act in a manner that is derived from moral principles. In this section we examine the methods by which these principles, in our case constraints on behavior derived from the LOW and ROE, can be represented effectively within a computational agent. We first focus on deontic logics as a primary source for implementation, then consider and dismiss utilitarian models, and bypass virtue ethics entirely (e.g., [Coleman 01]) as it does not lend itself well by definition to a model based on a strict ethical code.

Deontic Logics

Modal logics, rather than standard formal logics, provide a framework for distinguishing between what is permitted and what is required [Moor 06]. For ethical reasoning this clearly has pragmatic importance, and is used by a number of research groups worldwide in support of computational ethics. Moor observes that deontic logic (for obligations and permissions), epistemic logic (for beliefs and knowledge) and action logic (for actions) all can have a role "that could describe ethical situations with sufficient precision to make ethical judgments by a machine". A description of the operation of deontic logic is well beyond the scope of this paper; the reader is referred to [Horty 01] for a detailed exposition.

A research group at RPI [Bringsjord et al. 06] is quite optimistic about the use of deontic logic as a basis for producing ethical behavior in intelligent robots for three reasons:

1. Logic has been used for millennia by ethicists.
2. Logic and artificial intelligence have been very successful partners and computer science arose from logic.
3. The use of mechanized formal proofs with their ability to explain how a conclusion was arrived at is central for establishing trust.

They [Arkoudas et al. 05] argue for the use of standard deontic logics for building ethical robots, to provide proofs that <u>(1) a robot take only permissible actions</u> and (2) that <u>obligatory actions are indeed performed</u>, subject to ties and conflicts among available actions. They use the Athena interactive theorem proving framework for their work. This approach seems useful for more general ethical behavior with complex nuances, but has yet to be considered in a real-time application. However, the ROE and LOW have already been distilled from ethical first principles by people and may not require the complex reasoning methods used in their work. The robotic agent must only abide by them, not derive them.

They further insist that for a robot to be certifiably ethical, every meaningful action must access a proof that the action is at least permissible. This form of reasoning is quite consistent with the formalisms that were developed in Section 3.2. Outstanding questions remain regarding real-time computation for a computationally constrained agent. They argue this is feasible by using methods that encode back to first order logic and claim that even dealing with formulas as numerous as 4 million they can reason over these sets "sufficiently fast". Coupled with continuing advances in computational speed along the lines of Moore's Law their claims appear plausible. This is a strong candidate for implementation in our research.

The ethical code $\boldsymbol{C}$ a robot uses is not bound to any particular ethical theory. It can be deontological, utilitarian or whatever, according to [Bringsjord et al. 06]. The concepts of prohibition, permissibility, and obligation are central to deontic logics. The formalization of $\boldsymbol{C}$ in a particular computational logic $\mathbf{L}$ is represented as $\mathbf{\Phi}_{\boldsymbol{C}}^{\mathbf{L}}$. This basically reduces the problem for our ethical governor to the need to derive from the LOW and ROE a suitable $\mathbf{\Phi}^{\mathbf{L}}_{\mathbf{LOW} \cup \mathbf{ROE}}$, with the leading candidate for $\mathbf{L}$ being a form of deontic logic. Accompanying this ethical formalization is an ethics-free ontology which represents the core concepts that $\boldsymbol{C}$ presupposes (structures for time, events, actions, agents, etc.). A signature is developed that encodes the ontological concepts with special predicate letters and functions. Clearly this is an action item for our research, if deontic logic is to be employed in the use of lethality for ethical systems. There is much more involved as outlined in [Bringsjord et al. 06] but a pathway for the development of such a system seems feasible using these methods. They show one example using a variation of Horty's multi-agent deontic logic [Horty 01] applied to ethics in a medical domain using the Athena framework [Arkoudas et al. 05]. Both Athena and this version of Horty's logic should be given serious consideration for application within the ethical architecture. A first order of business in the near-term development of the ethical governor is the generation of an example using these tools and techniques that spans a limited space of warfare situations, using the ethical scenarios presented in Section 6.

Another research group that uses deontic logic for ethical reasoning [Wiegel 06, Van Den Hove et al. 02] couples the use of the well-know BDI model (belief-desire-intention) with a deontic-epistemic-action logic (DEAL) to model and specify activities for moral agents. Wiegel describes a set of design principles that are requirements for an artificial ethical system in a general sense. The relevant requirements that apply to a lethal autonomous agent bound by the LOW and ROE include:

- Bounded rationality, time and resource constraints
- Mixed moral and non-moral activity support and goals support
- Extendibility, formality, scalability, comprehensibility, and configurability

Regarding design principles, Wiegel advocates (presented with my comments regarding their relevance to our work):

1. Agents are proactive, goal-driven and reactive (consistent with a hybrid deliberative-reactive architecture)
2. Behavior is built from small action components (compatible with behavior-based robotic design)
3. Agents can decide if and when to update their information base (somewhat analogous to our ethical adaptor function)
4. Agents interact with each other and the environment (a given for an autonomous robotic system)

The DEAL Framework refers to a deontic component (right, obligation, permission, or duty) of an action on an epistemic (component of knowledge). It supplements deontic logic (that uses the basic operator $\boldsymbol{O}$ "it is obliged that") with epistemic logic that incorporates assertions about knowing and believing, and action logic that includes an action operator referred to as STIT (See To It That) [Van den Hoven 02].A typical assertion would be:

$$\mathbf{B_i\ (G(\Phi)) \rightarrow \boldsymbol{O}([i\ STIT\ \Phi])}$$

which asserts that if $\mathbf{i}$ believes that $\mathbf{\Phi}$ is good, then it should act in a manner to see that $\mathbf{\Phi}$ occurs. Roles and rights and the obligations associated with those rights are represented as a matrix. The obligations are actions defined using specific instances of the **STIT** operator. The agent's desires form intentions that trigger the ethical reasoning process.

Wiegel states this method constitutes a specification language rather than a formal language capable of theorem proving. This framework is implemented in an agent-oriented manner using the Java-based JACK agent language and development environment. [Wiegel et al. 05] presents the details of the implementation. They contend that the computational complexities are comparable to first-order predicate logic.

An interesting concept of potential relevance to our research is their introduction of the notion of a trigger, which invokes the necessary ethical reasoning at an appropriate time. In our case, the trigger for the use of the moral component of the autonomous system architecture would be the presence of a potential lethal action, a much more recognizable form of a need for an ethical evaluation, than for a more general setting such as business or medical practice. The mere presence of an active lethal behavior is a sufficient condition to invoke ethical reasoning.

[Wiegel 04] provides several useful lessons learned that may be of value for the implementation of an ethical governor:

1. Negative moral commands (obligations) are difficult to implement. Agents must be able to evaluate the outcomes of their actions, and classify them as right or wrong.
2. Morality must act as both a restraint and goal-director. In our case this is straightforward by virtue of the problem domain.
3. Restricting the amount of information may be required to avoid an agent being prevented from making a decision. This can be handled in our case by always reserving the right not to fire unless a properly informed decision has been made.
4. Moral epistemology is the major challenge. Typing of perceptions, events, facts, etc., have to be done at design-time.

Utilitarian Methods

Utilitarianism at first blush offers an appeal due to its ease of implementation as it utilizes a formal mathematical calculus to determine what the best ethical action is at any given time, typically by computing the maximum goodness (however defined) over all of the actors involved in the decision. [Anderson et al. 04] implemented an ethical reasoning system called Jeremy that is capable of conducting moral arithmetic. It is based on Bentham's Hedonistic Act Utilitarianism. The classical formulation of utilitarianism is to choose an action that maximizes good, pleasure or happiness over all of the parties involved. Jeremy uses pleasure and displeasure for its computational basis, simply adding the total pleasure for all individuals then subtracting the total displeasure for all to yield the total net pleasure. The values are determined from the product of the intensity, duration, and probability of their occurrence. The action selected is the one that provides the greatest total net pleasure. If a tie occurs, either action is considered equally correct. An integer is provided by a user of the system to quantify the pleasure in the range [-2,+2], with the likelihood of their occurrence chosen from {0.8,0.5,0.2} and other values for parameters of intensity. While this method is of academic interest, Utilitarian methods in general, do not protect the fundamental rights of an individual (e.g., a noncombatant) and are thus inappropriate for our goals at the highest level.

The Utilibot project [Cloos 05] was proposed as a system that uses act utilitarianism to maximize human well-being in the case of a hybrid health care/service robot for home use. It was intended to be implemented within a hybrid deliberative-reactive architecture, as is the case for this project. Subsequent to the original paper, however, no further reports were encountered, so one can only speculate if any results were obtained and what the specific technical details were.

[Grau 06] also dismisses the use of a utilitarian theory as a basis for a project such as outlined in this article, concluding: "Developing a utilitarian robot might be a reasonable project – even though the robot shouldn't treat humans along utilitarian lines and it wouldn't be a suitable ethical advisor for humans". I agree with his conclusions regarding its limited applicability and we will use other approaches as the basis for an ethical autonomous system capable of lethality.

Kantian Rule-based Methods

[Powers 06] advocates the use of rules for machine ethics: "A rule-based ethical theory is a good candidate for the practical reasoning of machine ethics because it generates duties or rules for action, and rules are (for the most part) computationally tractable." Indeed, computational tractability is a concern for logic-based methods in general. Powers states that Kant's categorical imperative lends itself to a rule-based implementation. This high-level principle, that forms the basis for a deontological school of ethical thought, is relatively vague when compared to the specific requirements for the ethical use of force as stated in the LOW and ROE. [Powers 05] lets the machine derive is own ethical theory which then can map prospective actions onto the deontic categories of forbidden, permissible, and obligatory. Maxims (a form of universal rules) are used to provide a consistency check for a suggested action. As an example, the machine might create the following Universals:

1. $\forall z \, \exists x \, \exists y \, (Cx \wedge Py) \rightarrow Az$ *A is obligatory for z*
2. $\forall z \, \exists x \, \exists y \, (Cx \wedge Py) \rightarrow \neg Az$ *A is forbidden for z*
3. $\neg\forall z \, \exists x \, \exists y \, (Cx \wedge Py) \rightarrow Az$ and $\neg\forall z \, \exists x \, \exists y \, (Cx \wedge Py) \rightarrow \neg Az$ *A is permissible for z*

where Cx=x is a circumstance, Py=y is a purpose and Az=z commits action A.

In our application, however, the LOW has effectively transformed the categorical imperative into a set of more direct and relevant assertions regarding acceptable actions towards noncombatants and their underlying rights, and the need for generalization by the autonomous system seems unnecessary. We need not have the machine derive its ethical rules on its own, so this approach is not relevant to our work.

4.2.2 *Particularism* - Case-based Reasoning

Generalism, as just discussed, appears appropriate for ethical reasoning based on the principles extracted from the LOW and ROE, but it may be less suitable for addressing responsibility attribution. [Johnstone 07] observes "There are however reasons to doubt whether this kind of analysis based on discrete actions and identifiable agents and outcomes, essentially, the attribution of responsibility, is adequate ….". We now investigate methods that may be particularly suitable for the responsibility advisor component of the ethical autonomous architecture under development.

McLaren used case-based reasoning (CBR) as a means of implementing an ethical reasoner [McLaren 06]. As our laboratory has considerable experience in the use of CBR for robotic control in robotic architectures ranging from reactive control [Ram et al. 97, Kira and Arkin 04, Likhachev et al. 02, Lee et al. 02] to deliberative aspects [Endo et al. 04, Ulam et al. 07] in a hybrid architecture, this method warrants consideration. Principles can be operationalized or extensionally defined, according to [McLaren 03], by directly linking them to facts represented in cases derived from previous experience.

McLaren has developed two systems implementing CBR, Truth-Teller and SIROCCO, both of which retrieve analogically relevant cases to the current situation. Unlike the previous ethical reasoning systems discussed, these do not arrive at an ethical decision, as he believes "reaching an ethical conclusion, in the end is a human decision maker's obligation" [McLaren 06]. Thus his system serves more as an ethical guide or assistant as opposed to a controller or decision-maker. It does provide an illustration that cases derived from previous experience can be retrieved based on their ethical content.

SIROCCO is the more relevant system for our application. It was intended "to explore and analyze the relationship between general principles and facts of cases" [McLaren 05]. Its domain is that of engineering ethics. Although SIROCCO's methods are of little value for the control of real-time ethical decision-making as required for the ethical governor and ethical behavioral control components of our architecture, its methods hold some promise for the responsibility advisor component as it is capable of making ethical suggestions drawn from experience to guide a user. It is an interpretive case-based reasoning system that can retrieve past cases and predict ethical codes that are relevant to the situation at hand.

The control flow of SIROCCO is shown in Figure 6. The mathematical details for surface retrieval and structural mapping appear in [McLaren 03]. In addition to the cases, the ethical codes in SIROCCO are represented as an action/event hierarchy, which characterizes the most important actions and events in ethics scenarios.

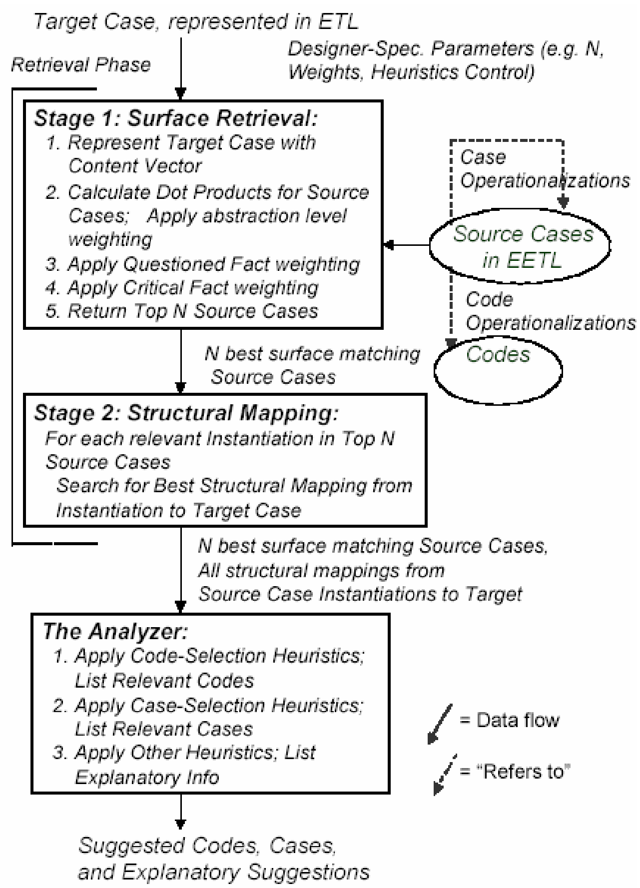


Figure 6: SIROCCO's Architecture (from [McLaren 05])

Case-based Reasoning has also been widely applied in the legal domain, and as the legal basis for the Laws of War define responsibility, no doubt additional insights can be gleaned from that research community.

An alternative CBR-based approach using a duty-based system was developed by [Anderson et al. 06] that *does* arrive at ethical conclusions derived from case data. W.D. is based on W.D. Ross's seven prima facie duties (establishing the ethical criteria) which combine Kantian duties and utilitarian principles with Rawls' theory of reflective equilibrium to provide a mechanism for reasoning over those criteria and arrive at an ethical decision [McLaren 07]. Rules (principles) are derived from cases provided by an expert ethicist who serves as a trainer. These rules are generalized as appropriate.

Horn Clause rules are derived from each training case using inductive logic programming, converging towards an equilibrium steady-state condition, where no further learning is required. From a representational perspective, a Horn Clause is a specific class of first order logic sentences that permit polynomial inference [Russell and Norvig 95]. The Prolog programming language is based upon this form of representation. Horn Clauses consist of assertions of the form:

$$P_1 \wedge P_2 \wedge \ldots P_n \Rightarrow Q$$

where P_i are nonnegated atoms. Figure 7 presents the learning algorithm used in W.D.

```
Input case and store in casebase
If case is covered by background knowledge or current hypothesis and its negative is not covered
        Then output correct action(s)
        Else
                Initialize list of case (PositiveCases) to contain all positive cases input so far
                Initialize list of cases (NegativeCases) to contain all negative cases input so far
                Initialize list of candidate clauses (CandClauses) to contain the clauses of current
                    hypothesis followed by an empty clause
                Initialize list of new hypothesis clauses (NewHyp) to empty list
                Repeat
                        Remove first clause (CurrentClause) from CandClauses
                        If CurrentClause covers a negative case in NegativeCases then
                                Generate all least specific specializations of CurrentClause and add
                                   those that cover a positive example in PositiveCases and not already
                                   present to CandClauses
                        Else add CurrentClause to NewHyp and remove all cases it covers from
                             PositiveCases
                Until PositiveCases is empty
                New hypothesis is the disjunction of all clauses in NewHyp
```

Figure 7: W.D.'s Inductive Logic Program Algorithm (from [Anderson et al. 05])

[Andersen et al. 05] developed a similar system, MedEthEx, for use in the medical ethics domain to serve as an advisor.

The end result for W.D. is the extraction of ethical rules from cases developed by expert trainers. This system seems well suited for learning ethics, but not necessarily for enforcing an already existing ethical standard, such as the LOW and ROE that we are concerned with for the run-time component of the architecture. The LOW and ROE directly provide a basis for representing the ethical rules without the limitations and dangers of training, and it is expected that logical assertions (Horn Clause or otherwise) will be generated to span this ethical space for autonomous lethal use in the battlefield. While the CBR method appears unsuitable for the run-time needs of the ethical governor or ethical behavioral control components, it may have value for the responsibility advisor, in terms of recalling experts' opinions on similar cases when deploying an autonomous lethal agent, and by making recommendations regarding responsibility to the operator accordingly. I agree with McLaren in principle regarding personal responsibility, where the onus lies on the human to make the decision and assume responsibility for the if, when, and how to use a lethal autonomous system (as is the case for any weapon system for that matter). Advice prior to deployment generated by a CBR system may be invaluable in assisting the person making that decision as it can be derived from expert ethicist's knowledge.

[Guarini 06] removes case-based methods a step further from reasoning directly from moral principles, by using a neural network to provide for the classification of moral cases. He thus avoids the use of principles entirely. Transparency is also lost as the system cannot justify its decisions in any meaningful way; i.e., explanations and arguments are not capable of being generated. For these reasons, this method offers considerably less utility in our application of bounded morality for the application of well-articulated ethical principles in the LOW and ROE. It will not be considered further here.

4.2.3 Ethical Decision-making

To help guide our decisions regarding representational content and implementation, it is useful to consider how soldiers are trained to apply the ethical decision-making method as a commander, leader or staff member [ATSC 07]. From this Army Training manual, the algorithm is specified as follows:

Performance Steps:

1. Clearly define the ethical problem.
2. Employ applicable laws and regulations.
3. Reflect on the ethical values and their ramifications.
4. Consider other applicable moral principles.
5. Reflect upon appropriate ethical theories.
6. Commit to and implement the best ethical solution.
7. Assess results and modify plan as required.

This may not be suitable for real-time decision making, as consideration and reflection are part of deliberation, which in the battlefield the soldier may not have the luxury of time to undertake.

Ultimately, a robotic system, however, will be able to compute more effectively over larger bodies of knowledge in shorter time periods than a human can.

It has been stated that "many, if not most, senior officers lean toward utilitarianism", which is interpreted as "Choose the greater (or greatest) good" [Toner 03]. Utilitarianism is recognized as an ethical framework that is capable of ignoring fundamental rights, which can be a serious flaw for this sort of battlefield bottom-line thinking from a legal perspective, where the ends are used to justify the means in lieu or preserving the law-given rights of noncombatants.

Recommendations for ethical decision-making are further refined in the United States Army Soldier's Guide (reproduced from [U.S. Army 04]) with Step 3a ensuring compliance with International Law:

The Ethical Reasoning Process

Step 1. Problem definition. Same as the problem solving steps.

Step 2. Know the relevant rules and values at stake. Laws, Army Regulations (ARs), ROE, command policies, Army values, etc.

Step 3. Develop possible courses of action (COA) and evaluate them using these criteria:

> a. Rules—Does the COA violate rules, laws, regulations, etc.? For example, torturing a prisoner might get him to reveal useful information that will save lives, but the law of war prohibits torture under any circumstances. Such a COA violates an absolute prohibition.
>
> b. Effects—After visualizing the effects of the COA, do you foresee bad effects that outweigh the good effects? For example, you are driving along a railroad and you see a train on the tracks. If you speed up to beat the train to the crossing, you might save a little time getting to your destination. But the potential bad effects outweigh the time you might save.
>
> c. Circumstances— Do the circumstances of the situation favor one of the values or rules in conflict? For example, your battle-buddy was at PT formation this morning but now is absent at work call formation. Do you cover for him? Your honor and loyalty to the unit outweigh your friendship and loyalty to your buddy, so the ethical COA would be to report the truth rather than lie about his whereabouts.
>
> d. "Gut check"—Does the COA "feel" like it is the right thing to do? Does it uphold Army values and develop your character or virtue? For example, you come upon a traffic accident and a number of vehicles have stopped, apparently to render aid, but you aren't sure. Stopping may cause further congestion in the area, but ensuring injured are cared for and that emergency services are on the way further strengthens the values of duty and honor.

Step 4. Now you should have at least one COA that has passed Step 3. If there is more than one COA, choose the course of action that is best aligned with the criteria in Step 3.

There is a clear mix of deontological methods (rights based upon the LOW in step 3a), followed by a utilitarian analysis in step 3b. Step 3d is clearly outside the scope of autonomous systems.

Another ethical analysis example specific to the military targeting process is outlined in [Bring 02]:

Those who plan or decide upon an attack shall:

1. Do everything feasible to verify that the objectives to be attacked are military objectives.
2. Take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event minimizing, incidental loss of civilian life.

3. Refrain from deciding to launch an attack that may be expected to cause such incidental loss, which would be excessive in relation to the concrete and direct military advantage anticipated.
4. Suspend an attack if it becomes apparent that it may be expected to cause incidental loss of civilian life, damage to civilian objects, or a combination thereof, "which would be excessive in relation to the concrete and direct military advantage anticipated."
5. In addition, "effective advance warning shall be given of attacks which may affect the civilian population, unless circumstances do not permit."

The inclusion of suspending an attack under certain conditions (step 4) is particularly relevant, which requires ongoing monitoring and feedback during the attack. This must be accommodated into the ethical architecture.

James Baker [Baker 02] describes the process by which he reviewed specific targets in the Kosovo campaign:

1. What is the military objective?
2. Are there collateral consequences?
3. Have we taken all appropriate measures to minimize those consequences and to discriminate between military objectives and civilian objects?
4. Does the target brief quickly and clearly identify the issues for the president and principals?

Items 1-3 clearly conform to the LOW. Item 4, however, seems a somewhat unusual criterion, but only adds more restrictions above and beyond the LOW. Of note is his comment regarding the tensions associated with what he refers to as "going downtown", a form of shock-and-awe strategy intended to bring a rapid end to the conflict, and "dual-use" targets for objects which support both military and civilian needs. The LOW appear to have been followed at least in this conflict, with Baker stating "Nor, I should be clear, am I suggesting the United States applied anything other than a strict test of military objective as recognized in customary international law …" [Baker 02]. He does state that these legal areas (e.g., dual use and shock-and-awe) warrant further review as they will be an issue in the future. He was correct.

From a non-military computational ethics perspective, [Maner et al. 02] surveyed a broad range of heuristic ethical decision-making processes in the literature (60), which he distilled into a series of stages, where some consensus existed on the correct procedure to achieve an ethical sound decision. The stages are:

1. *Preparing*: Develop and cultivate morality in the agent.
2. *Inspecting*: Look at the current situation and assess what is factually relevant (not just morally relevant).
3. *Elucidating*: Determine what is missing, then find it or make assumptions that cover the missing pieces. Clarify additional concepts, identify obstacles, and determine the affected

parties. Ask, “Should X do Y given Z?” and gather the information to answer such questions.

4. *Ascribing*: Infer values, goals, ideals, priorities, and motives and ascribe them to the parties involved.
5. *Optioning:* Brainstorm all possible courses of action available. Eliminate non-feasible ones.
6. *Predicting*: For each remaining option, list possible consequences, both long and short-term. Associate the risks and benefits with each participant.
7. *Focusing:* Determine who is sufficiently affected to be considered stakeholders among all affected parties. Determine rights, responsibilities, and duties. Determine which facts are morally relevant. Ask "Should X do or not do Y assuming Z?”.
8. *Calculating:* Use formal decision-making procedures to quantify impacts.
9. *Applying:* Consider each stakeholder/action pair separately. Catalog and rank reasons. Recognize which moral actions are required from those that are permitted but not required. Review laws, policies, and codes for parallels.
10. *Selecting*: Chose an option, confirm with common-sense ethical tests.
11. *Acting*: Plan how to carry out the action, and undertake it.
12. *Reflecting*: Monitor the decision and assess its consequences on the stakeholders. If needs be reconsider a better course of action in the future.

Although this model is concerned with social, professional, and business ethical decision-making and not the lethal force application that involves bounded morality and the rigid prescribed codes that we are concerned with, aspects of the procedure, apart from its utilitarian flavor, have value for developing a suitable ethical algorithm in our research (e.g., inspecting, elucidating, predicting, applying, selecting, acting, and reflecting).

Maner notes that many ethical procedures have serious limitations, including several which we should deliberately try to avoid in the design of the lethal agent architecture, including:

- An inability to deal with situations that change rapidly while under analysis
- The ethical issue is defined too early in the process
- They do not degrade gracefully under time pressure
- It may require a high-level of situational ethical awareness in the very first step
- Computational complexity problems in complex situations
- It may not allow a fact or assumption to be withdrawn once introduced

An interesting approach regarding the time pressure issues mentioned above, may involve the use of *anytime algorithms*, which start the reasoning regarding lethality from the most conservative stance and then progressively, as more justifications and obligations arrive, move closer towards lethal action. This reserves lethal force as a recourse of confirmed obligation. We revisit this in Section 5 when we discuss architectural implementations.

5. Architectural Considerations

We now move closer towards an implementation of the underlying theory developed in Section 3, using, as appropriate, the content and format of the representational knowledge described in Section 4. This is a challenging task, as deciding how to apply lethal force ethically is a difficult problem for people, let alone machines:

> *Whether deployed as peacekeepers, counterinsurgents, peace enforcers, or conventional warriors, United States ground troops sometimes make poor decisions about whether to fire their weapons. Far from justifying criticism of individual soldiers at the trigger, this fact provides the proper focus for systemic improvements. The problem arises when the soldier, having been placed where the use of deadly force may be necessary, encounters something and fails to assess correctly whether it is a threat. Then the soldier either shoots someone who posed no such threat, or surrenders some tactical advantage. The lost advantage may even permit a hostile element to kill the soldier or a comrade.* [Martins 94, p. 10]

Sometimes failure occurs because restraint is lacking (e.g., killing of unarmed civilians in My Lai in March 1968; Somalia in February 1993; Haditha in November 2005), in other cases it is due to the lack of initiative (e.g., Beirut truck bombing of Marine barracks, October 1983) [Martins 94]. Martins observes that unduly inhibited Soldiers, too reluctant to fire their weapons, prevent military units from achieving their objectives. In WWII most infantrymen never fired their weapons, including those with clear targets. Soldiers who fire too readily also erect obstacles to tactical and strategic success. We must strike a delicate balance between the ability to effectively execute mission objectives with the absolute compliance that the Laws of War will be observed.

To address these problems, normally we would turn to neuroscience and psychology to assist in the determination of an architecture capable of ethical reasoning. This paradigm has worked well in the past [Arkin 89, Arkin 92, Arkin 05]. Relatively little is known, however, about the specific processing of morality by the brain from an architectural perspective or how this form of ethical reasoning intervenes in the production and control of behavior, although some recent advances in understanding are emerging [Moll et al. 05, Tancredi 05]. [Gazzaniga 05] states: “Abstract moral reasoning, brain imaging is showing us, uses many brain systems”. He identifies three aspects of moral cognition:

1. Moral emotions which are centered in the brainstem and limbic system.
2. Theory of mind, which enables us to judge how others both act and interpret our actions to guide our social behavior, where mirror neurons, the medial structure of the amygdala, and the superior temporal sulcus are implicated in this activity.
3. Abstract moral reasoning, which uses many different components of the brain.

Gazzaniga postulates that moral ideas are generated by an interpreter located in the left hemisphere of our brain that creates and supports beliefs. Although this may be useful for providing an understanding for the basis of human moral decisions, it provides little insight into the question that we are most interested in, i.e., how, once a moral stance is taken, is that enforced upon an underlying architecture or control system. The robot need not derive the underlying moral precepts; it needs solely to apply them. Especially in the case of a battlefield robot (but also for a human soldier), we do not want the agent to be able to derive its own beliefs regarding the moral implications of the use of lethal force, but rather to be able to apply those that have been previously derived by humanity as prescribed in the LOW and ROE.

[Hauser 06] argues that "all humans are endowed with a moral faculty – a capacity that enables each individual to unconsciously and automatically evaluate a limitless variety of actions in terms of principles that dictate what is permissible, obligatory, or forbidden", attributing the origin of these ideas to Adam Smith and David Hume. When left at this descriptive level, it provides little value for an implementation in an autonomous system. He goes a step further, however, postulating a *universal moral grammar* of action that parallels Chomsky's generative grammars for linguistics, where each different culture expresses its own set of morals, but the nature of the grammar itself restricts the overall possible variation, so at once it is both universal and specific. This grammar can be used to judge whether actions are permissible, obligatory, or forbidden. Hauser specifies that this grammar operates without conscious reasoning, but more importantly without explicit access to the underlying principles, and for this reason may have little relevance to our research. The principles (LOW) we are dealing with are explicit and not necessarily intuitive.

Nonetheless, Hauser (p. 31) also observes that ethical decisions are based on different architectural "design specs", which seem to closely coincide with the reactive/deliberative partitioning found in hybrid autonomous system architectures [Arkin 98]. His first ethical system model is based upon *intuitions* (Humean) which are "fast, automatic, involuntary, require little attention, appear early in development, are delivered in the absence of principled reasons, and often appear immune to counter-reasoning". The second design is *principled reasoning* (Kantian) which is "slow, deliberate, thoughtful, justifiable, requires considerable attention, appears late in development, justificable, and open to carefully defended and principled counterclaims". This division creates opportunities for introducing ethical decision-making at both the deliberative and reactive components of a robotic architecture, which will be explored further in this section, albeit using different approaches. Hauser identifies three different architectural models, shown in Figure 8, which can potentially influence the design of an ethical autonomous system.

He further contends that the third model (shown in Figure 8C) is the basis for human ethical reasoning, which is based on earlier work by Rawls and supported by recent additional neuroimaging evidence. From my reading of [Rawls 71, Ch. 2], however, and the principles of justice that he provides as an alternative to utilitarianism, it is unclear how this is connected to the more immediate and intuitive action analysis that Hauser describes as the basis for his third architectural model. But no matter, Hauser's Rawlsian model is based more on human intuitions rather than on formal rules and laws (e.g., LOW) as will be required for our particular application for an ethical basis of lethality in autonomous systems. Nor is it particularly relevant that the same models of ethical reasoning that are postulated for humans be applied to battlefield robots, especially given the typical failings of humanity under these extremely adverse conditions. Instead, importing a variant of the model shown in Figure 8B seems a more appropriate and relatively straightforward approach to implement within an existing deliberative/reactive architecture [Arkin and Balch 97], since many machine ethical systems utilize logical reasoning methods (deontological or utilitarian) that are suitable for a modular moral faculty component. In addition, expanded models of our existing methods for affective control [Arkin 05] can be utilized in our system as part of an ethical adaptor component. The focus for the reactive ethical architectural component for ethical behavioral control will not involve emotion directly, however, as that has been shown to impede the ethical judgment of humans in wartime [Surgeon General 06].

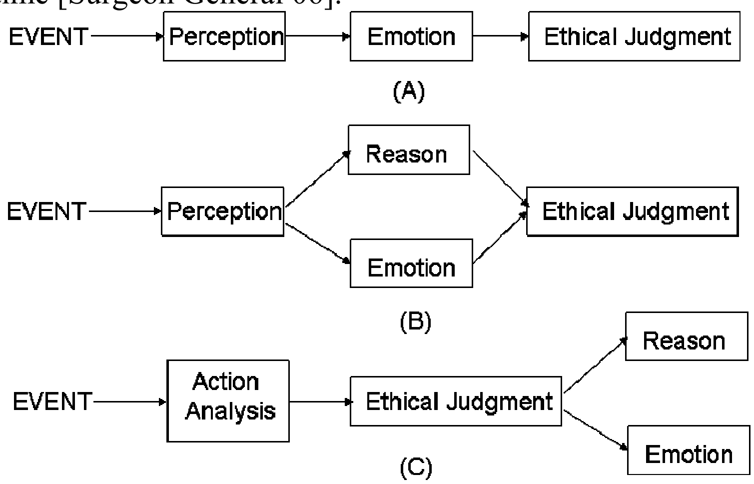


Figure 8: Three Human Ethical Architectural Candidates [Hauser 06]

(A) Corresponds to Hume's view: emotion determines the ethical judgment.
(B) Hybrid Kantian/Humean architecture: both reason and emotion determine ethical judgment.
(C) Rawlsian architecture: action analysis in itself determines the ethical judgment and emotion and reason follow post facto.

5.2 Architectural Requirements

In several respects, the design of an autonomous system capable of lethal force can be considered as not simply an ethical issue, but also a safety issue, where safety extends to friendly-force combatants, noncombatants, and non-military objects. The Department of Defense is already developing an unmanned systems safety guide for acquisition purposes [DOD 07]. Identified safety concerns not only include the inadvertent or erroneous firing of weapons, but the potentially ethical question of erroneous target identification that can result in a mishap of engagement of, or firing upon, unintended targets. Design precept DSP-1 states that the Unmanned System shall be designed to minimize the mishap risk during all life cycle phases [DOD 07]. This implies that consideration of the LOW and ROE must be undertaken from the onset of the design of an autonomous weapon system, as that is what determines, to a high degree, what constitutes an unintended target.

Erroneous target identification occurs from poor discrimination, which is a consequence of inadequate situational awareness. Figure 9 illustrates the trend for autonomous situational awareness as the levels of autonomy increase. Situational awareness is defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the future" [DOD 07]. Note that the onset of autonomy is not discontinuous but rather follows a smooth curve, permitting a gradual introduction of capability into the battlefield as the technology progresses.

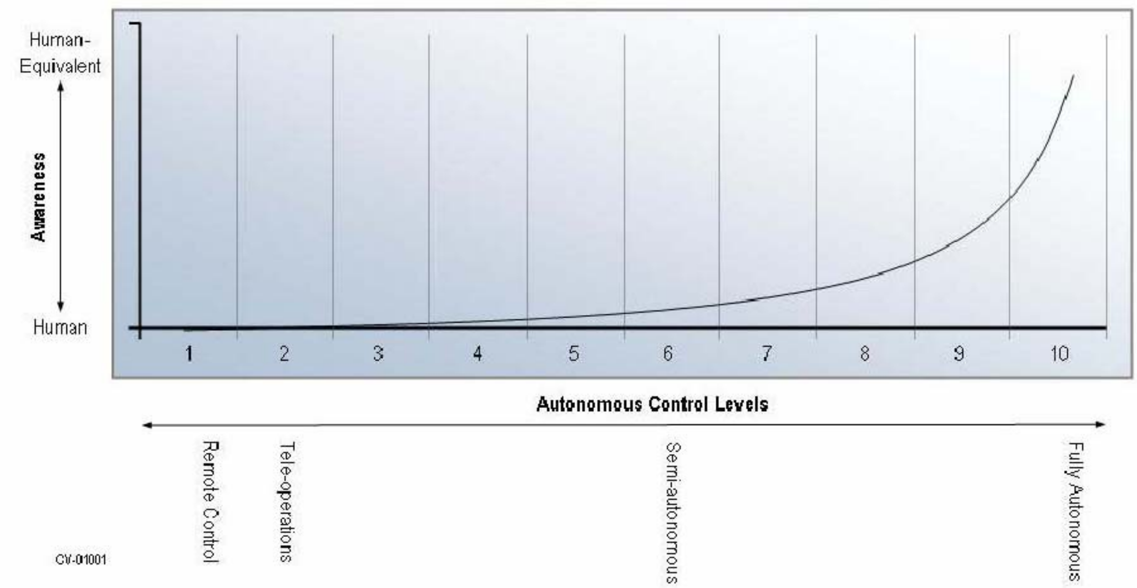


Figure 9: Illustration of the Increasing Requirement for Machine Situational Awareness as Autonomy Increases (source: [DOD 07]).

[Parks 02] listed a series of factors that can guide the requirements for appropriate situational awareness in support of target discrimination and proportionality. They are summarized in Figure 10.

Target intelligence	Distance to target	Target winds, weather
Planning time	Force training, experience	Effects of previous strikes
Force integrity	Weapon availability	Enemy defenses
Target identification	Target acquisition	Rules of engagement
Enemy intermingling	Human factor	Equipment failure
Fog of war		

Fig. 10: Factors Affecting Collateral Damage and Collateral Civilian Casualties [Parks 02]

It is a design goal of this project to be able to produce autonomous system performance that not only equals but exceeds human levels of capability in the battlefield from an ethical standpoint. How can higher ethical standards be achieved for an ethical autonomous systems than that of a human? Unfortunately, we have already observed in Section 1.1 there is plenty of room for improvement. Some possible answers are included in the architectural desiderata for this system:

1. Permission to kill alone is inadequate, the mission must explicitly obligate the use of lethal force.
2. The Principle of Double Intention, which extends beyond the LOW requirement for the Principle of Double Effect, is enforced.
3. In appropriate circumstances, novel tactics can be used by the robot to encourage surrender over lethal force, which is feasible due to the reduced or eliminated requirement of self-preservation for the autonomous system.

4. Strong evidence of hostility is required (fired upon or clear hostile intent), not simply the possession or display of a weapon. New robotic tactics can be developed to determine hostile intent without premature use of lethal force (e.g., close approach, inspection, or other methods to force the hand of a suspected combatant).
5. In dealing with POWs, the system possesses no lingering anger after surrender, thus reprisals are not possible.
6. There is never intent to deliberately target a noncombatant.
7. Proportionality may be more effectively determined given the absence of a strong requirement for self-preservation, reducing the need for overwhelming force.
8. Any system request to invoke a privileged response (lethality) automatically triggers an ethical evaluation.
9. Adhering to the principle of "first, do no harm", which indicates that in the absence of certainty (as defined by λ and τ) the system is forbidden from acting in a lethal manner. Perceptual classes (p,λ) and their associated τ should be defined appropriately to only permit lethality in cases where clear confirmation of a discriminated target is available and ideally supported by ideally multiple sources of evidence.

Considering our earlier discussion on forbidden and obligatory actions (Sec. 4), the architecture must also make provision for ensuring that forbidden lethal actions as specified by the LOW are not undertaken under any circumstances, and that lethal obligatory actions (as prescribed in the ROE) are conducted when not in conflict with LOW (as they should be). Simple permissibility for a lethal action is inadequate justification for the use of lethal force for an autonomous system. The LOW disables and the ROE enables the use of lethal action by an autonomous system.

The basic procedure underlying the overall ethical architectural components can be seen in Figure 11. It addresses the issues of responsibility, military necessity, target discrimination, proportionality, and the application of the Principle of Double Intention (acting in a way to minimize civilian collateral damage). Algorithmically:

```
Before acting with lethal force
        ASSIGN RESPONSIBILITY (A priori)
        ESTABLISH MILITARY NECESSITY
        MAXIMIZE DISCRIMINATION
        MINIMIZE FORCE REQUIRED (PROPORTIONALITY+DOUBLE INTENTION)
```

The architectural design is what must implement these processes effectively, efficiently, and consistent with the constraints derived from the LOW and ROE.

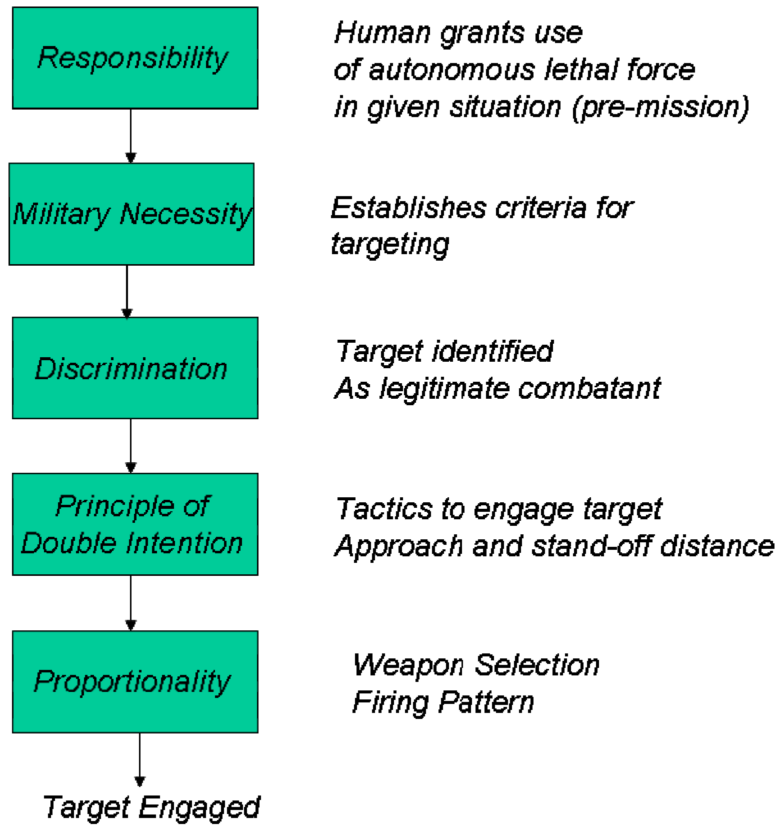


Figure 11: Ethical Architectural Principle and Procedure

This can be refined further into a set of additional requirements:

1. Discrimination
   a. Distinguish civilian from enemy combatant
   b. Distinguish enemy combatant from non-combatant (surrender)
   c. Direct force only against military objectives
2. Proportionality
   a. Use only lawful weapons
   b. Employ an appropriate level of force (requires the prediction of collateral damage and military advantage gained)
3. Adhere to Principle of Double Intention
   a. Act in a manner that minimizes collateral damage
   b. Self-defense does not justify/excuse the taking of civilian lives [Woodruff 82]
4. In order to fire, the following is required:

$$[\{\forall c_{\texttt{forbidden}} \mid c_{\texttt{forbidden}}(\mathbf{S_i})\} \wedge \{\exists c_{\texttt{obligate}} \mid c_{\texttt{obligate}}(\mathbf{S_i})\}] \Leftrightarrow \mathrm{PTF}(\mathbf{S_i})$$

for $c_{forbidden}, c_{obligate} \in C$, situation $\mathbf{S_i}$ and binary predicate **PTF** Permission-to-Fire. This clause states that in order to have permission to fire in this situation, all forbidden constraints must be upheld, and at least one obligating constraint must be true. **PTF** must be **TRUE** for the weapon systems to be engaged.

5. If operator overriding of the ethical governor's decision regarding permission to fire is allowed, we now have:

$$(\text{OVERRIDE}(\mathbf{S_i}) \text{ xor } [\{\forall c_{forbidden} \mid c_{forbidden}(\mathbf{S_i})\} \wedge \{\exists c_{obligate} \mid c_{obligate}(\mathbf{S_i})\}]) \Leftrightarrow \text{PTF}(\mathbf{S_i})$$

By providing this override capability, the autonomous system no longer maintains the right of refusal of an order, and ultimate authority vests with the operator. The logic and design recommendations underlying operator overrides are discussed in the Responsibility Advisor section (Sec. 5.2.4).

6. Determine the effect on mission planning (deliberative component's need to replan) in the event of an autonomous system's refusal to engage a target on ethical grounds.
7. Incorporate additional information from network-centric warfare resources as needed to support target discrimination. "Network Centric Warfare and Operations, fundamental tenets of future military operations, will only be possible with the Global Information Grid (GIG) serving as the primary enabler of critical information exchange." [DARPA 07]

Other miscellaneous information, that can be utilized within the architecture guidelines includes:

1. Regarding weapon tactics:
   - An argument is often made that "Shooting to wound is unrealistic and because of high miss rates and poor stopping effectiveness, can prove dangerous for the Marine and others." Nonetheless shoot to wound ROE may use language such as "when firing, shots will be aimed to wound, *if possible*, rather than kill" [CLAMO 02].
   - Warning shots may or may not be authorized depending on the applicable ROE for an operation.
2. Regarding battlefield carnage which is computed as the sum of

   (A) Intended Combatants +
   (B) Unintended Friendly forces (Fratricide) +
   (C) Intended non-combatants +
   (D) Unintended non-combatants (collateral)

   where:

   - (A) is intended and consistent with the LOW and determined by mission requirements (ROE).
   - (B) is unintended and inconsistent with ROE – minimize to 0 (i.e., eliminate accidental deaths).
   - (C) is intended but inconsistent with LOW – must be designed to be always 0 (i.e., removal of irrational unethical behavior)
   - (D) may or may not be acceptable given the LOW, the Principle of Double Effect, and the ROE. Apply the Principle of Double Intention to minimize collateral damage by adjusting proportionality as needed given military necessity.

Thus the design goal regarding battlefield carnage becomes to conduct (A) consistent with mission objectives, completely eliminate (B) and (C), and to minimize (D).

5.2 Architectural Design Options

We return now to the actual design of the overall system. Multiple architectural opportunities are presented below that can potentially integrate a moral faculty into a typical hybrid deliberative/reactive architecture [Arkin 98] (Fig. 12). These components are:

1. **Ethical Governor**: A transformer/suppressor of system-generated lethal action ($\boldsymbol{\rho}_{lethal\text{-}ij}$) to permissible action (either nonlethal or obligated ethical lethal force $\boldsymbol{\rho}_{l\text{-}ethicall\text{-}ij}$). This deliberate bottleneck is introduced into the architecture, in essence, to force a second opinion prior to the conduct of a privileged lethal behavioral response.
2. **Ethical Behavioral Control:** This design approach constrains all individual controller behaviors (β_i) to only be capable of producing lethal responses that fall within acceptable ethical bounds ($\mathbf{r}_{\mathbf{l\text{-}ethical\text{-}ij}}$).
3. **Ethical Adaptor:** This architectural component provides an ability to update the autonomous agent's constraint set (C) and ethically related behavioral parameters, but only in a more restrictive manner. It is based upon both an after-action reflective review of the system's performance or by using a set of affective functions (e.g., guilt, remorse, grief, etc.) that are produced if a violation of the LOW or ROE occurs.
4. **Responsibility Advisor**: This component forms a part of the human-robot interaction component used for pre-mission planning and managing operator overrides. It advises in advance of the mission, the operator(s) and commander(s) of their ethical responsibilities should the lethal autonomous system be deployed for a specific battlefield situation. It requires their explicit acceptance (authorization) prior to its use. It also informs them regarding any changes in the system configuration, especially in regards to the constraint set C. In addition, it requires operator responsibility acceptance in the event of a deliberate override of an ethical constraint preventing the autonomous agent from acting.

The preliminary specifications and design for each of these system components is described in more detail below. Note that these systems are intended to be fully compatible with each other, where the ideal overall design would incorporate all four of these architectural components. To a high degree, they can be developed and implemented independently, as long as they operate under a common constraint set C.

The value of clearly segregating ethical responsibility in autonomous systems has been noted by others. "As systems get more sophisticated and their ability to function autonomously in different context and environment expands, it will become important for them to have 'ethical subroutines' of their own… these machines must be self-governing, capable of assessing the ethical acceptability of the options they face" [Allen et al. 06]. The four architectural approaches advocated above embody that spirit, but they are considerably more complex than simple subroutines.

It must be recognized again, that this project represents a very early stage in the development of an ethical robotic architecture. Multiple difficult open questions remain that entire research programs can be crafted around. Some of these outstanding issues involve: the use of proactive tactics or intelligence to enhance target discrimination; recognition of a previously identified legitimate target as surrendered or wounded (change to POW status); fully automated combatant/noncombatant discrimination in battlefield conditions; proportionality optimization using the Principle of Double Intention over a given set of weapons systems and methods of employment; in-the-field assessment of military necessity; to name but a few. Strong (and limiting) simplifying assumptions will be made regarding the ultimate solvability of these problems in the discussions that follow, and as such this should temper any optimism involving the ability to field an ethical autonomous agent capable of lethality in the near term.

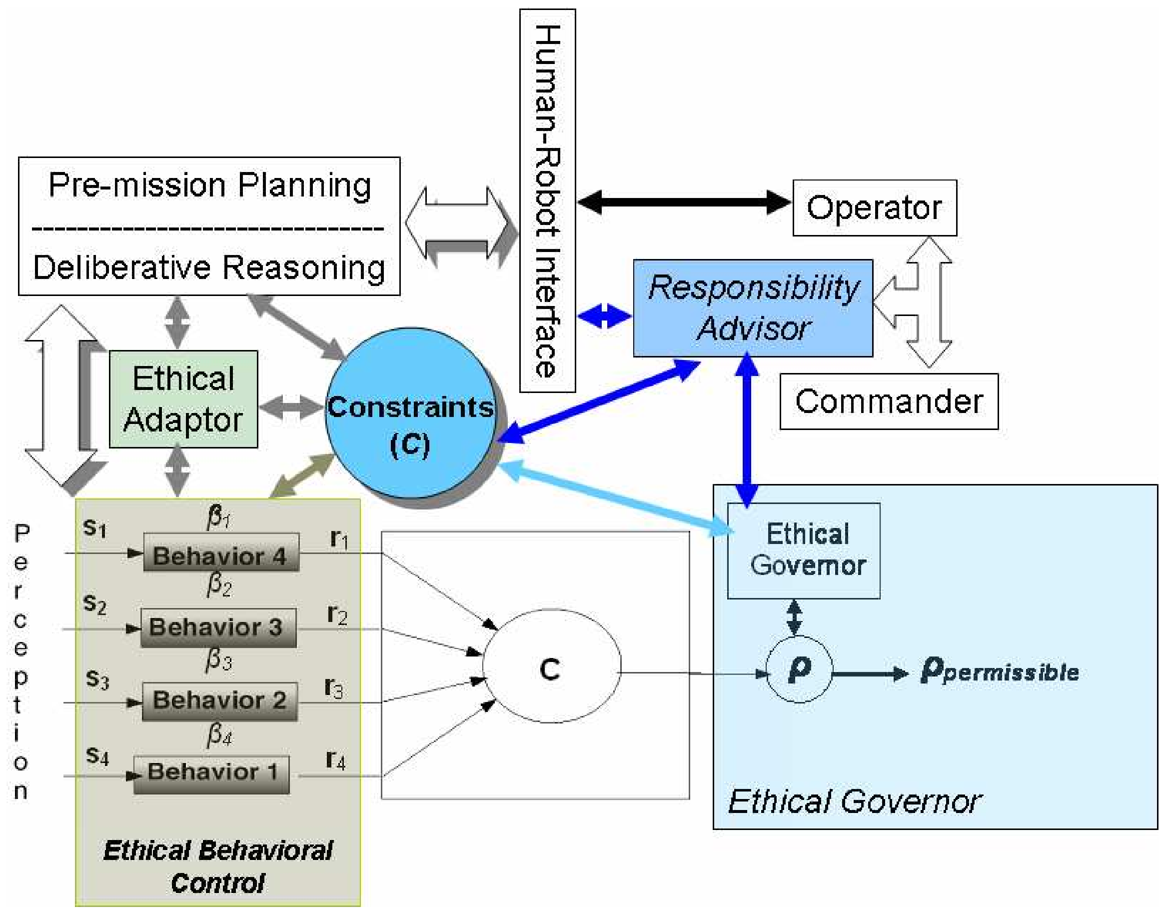


Figure 12: Major Components of an Ethical Autonomous Robot Architecture. The newly developed ethical components are shown in color.

5.2.1 Ethical Governor

This section outlines the design for the ethical governor component of the architecture. This component's responsibility is to conduct an evaluation of the ethical appropriateness of any lethal response that has been produced by the robot architecture prior to its being enacted. It can be largely viewed as a bolt-on component between the hybrid architectural system and the actuators, intervening as necessary to prevent an unethical response from occurring. Technically, the governor can be considered a part of the overall deliberative system of the architecture that is concerned with response evaluation and confirmation. It is considered a separate component, however, in this work as it does not require high-levels of interaction with the other main components of deliberation (although it can request replanning) and it can be deployed in an otherwise purely reactive architecture if desired.

The term governor is inspired by Watts' invention of the mechanical governor for the steam engine, a device that was intended to ensure that the mechanism behaved safely and within predefined bounds of performance. As the reactive component of a behavioral architecture is in essence a behavioral engine intended for robotic performance, the same notion applies, where here the performance bounds are ethical ones. Figure 13 illustrates this design and its relationship to Watts' original concept.

Recall that the overt robotic response $\boldsymbol{\rho} = \mathbf{C}(\mathbf{G} * \mathbf{B}(\mathbf{S_i}))$ is the behavioral response of the agent to a given situation $\mathbf{S_i}$. To ensure an ethical response, the following must hold:

$$\{\forall\, \boldsymbol{\rho} \mid \boldsymbol{\rho} \notin \boldsymbol{P}_{\textit{l-unethical}}\}$$

Formally, the role of the governor is to ensure that an overt lethal response $\boldsymbol{\rho}_{\textit{lethal-ij}}$ for a given situation is ethical, by confirming that it is either within the response set $\boldsymbol{P}_{\textit{l-ethical}}$ or is prevented from being executed by mapping an unethical $\boldsymbol{\rho}_{\textit{lethal-ij}}$ onto the null response $\varnothing$ (i.e., ensuring that it is ethically permissible). If the ethical governor needs to intervene, it must send a notification to the deliberative system in order to allow for replanning at either a tactical or mission level as appropriate, and to advise the operator of a potential ethical infraction of a constraint or constraints c_k in the ethical constraint set C.

Each constraint $c_k \in C$ specified must have at least the following data fields:

1. **Logical form**: As derived from deontic logic (Section 4.2.1). Horty's Deontic Logic is the current candidate of choice for this, possibly using tools and techniques from [Bringsjord et al. 06] or [Wiegel et al. 05].
2. **Textual descriptions:** Both a high-level and detailed description for use by the Responsibility Advisor.
3. **Active status flag**: Allowing for mission-relevant ROE to be defined within an existing set of constraints, and to designate operator overrides (Section 5.2.4).
4. **Base types**: Forbidden (e.g., LOW or ROE derived) or obligated (e.g., ROE derived). These will be relegated to either a long-term memory (LTM) for those constraints which

persist over all missions, or a short-term memory (STM) for those constraints that are derived from the specific current ROE for the given Operational Orders. Changes in LTM, that encode the LOW, require special two-key (Section 5.2.4) permission.

5. **Classification**: One chosen from Military Necessity, Proportionality, Discrimination, Principle of Double Intention, and Other. This is used only to facilitate processing by ordering the application of constraints by Class.

Other constraint fields may be added in the future as this research progresses.

Constraints are created and added to the system by the developer through the use of a graphical user interface (GUI) referred to as the constraint editor. It provides the means for filling out the necessary fields prior to their addition to the constraint set, as well as conducting accuracy checking and confirmation.

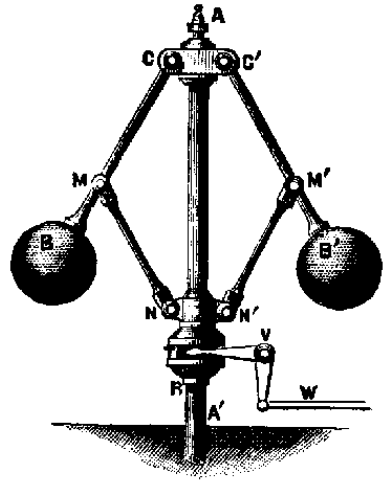
Control within the ethical governor is an open research question at the time of this writing, but several methods are expected to be used and are outlined below. Real-time control will need to be achieved for in-the-field reasoning. This assumes that the perceptual system of the architecture, charged with producing a certainty measure λ for each relevant stimulus (e.g., candidate target) $\mathbf{s} \in S$ that is represented as a binary tuple (p,λ), where p is a perceptual class (e.g., combatant or noncombatant). In addition, a mission-contextual threshold τ for each relevant perceptual class is also evaluated. Mission-specific thresholds are set prior to the onset of the operation, and it is expected that case-based reasoning (CBR) methods, as already employed in our Navy research on pre-mission planning for littoral operations for teams of UxVs, can effectively provide such system support [Ulam et al. 07]. Assessment of proportionality may also be feasible via the use of CBR by using previous weapons experience based on successful ethical practice as the basis for future action. Discrimination trees based on LOW may also serve as a method for legitimizing targets.

It is a major assumption of this research that accurate target discrimination with associated uncertainty measures can be achieved despite the fog of war, but it is believed that it is ultimately possible for the reasons as stated in Section 1.1. The architecture described herein is intended to provide a basis for ethically acting upon that information once produced.

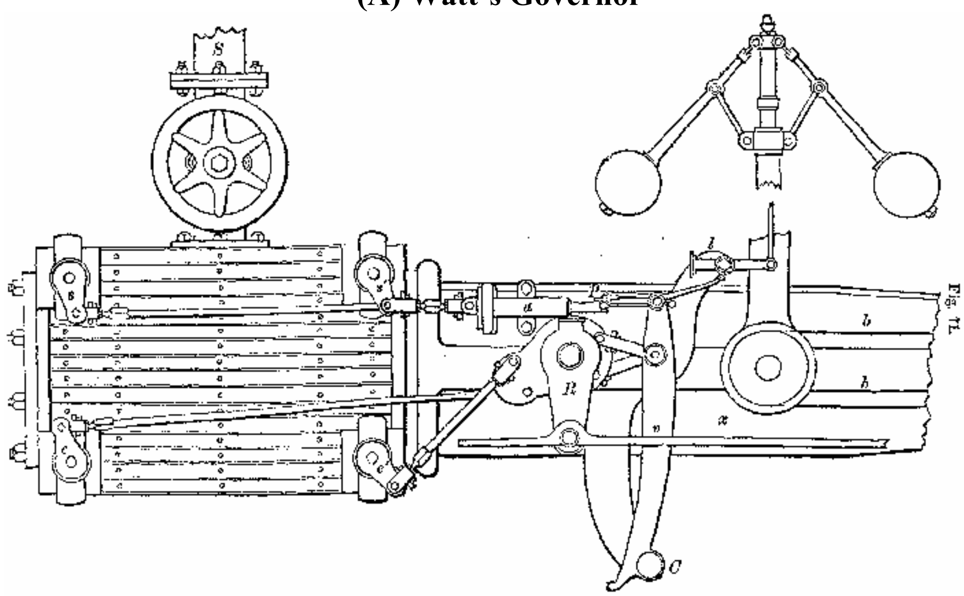
Given the ethical governor's real-time requirements, it is anticipated that an anytime algorithm [Zilberstein 96] will be required, always acting in the most conservative manner to ensure that the LOW is adhered to, while progressively migrating from a conservative to a more aggressive method as obligations are evaluated.

To achieve this level of performance, the ethical governor (Figure 14) will require inputs from:

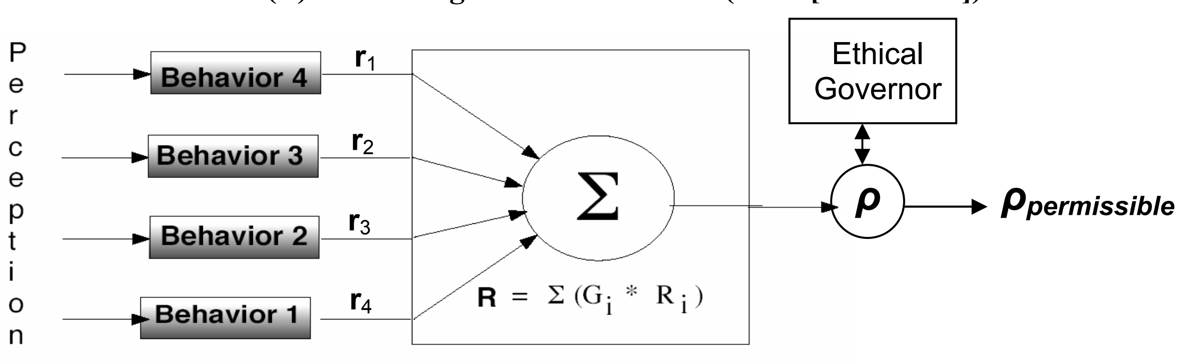
1. The overt response generated by the behavioral controller, $\boldsymbol{\rho}$
2. The perceptual system
3. The Constraint Set C (both long-term and short-term memory)
4. The Global Information Grid (GIG) to provide additional external sources of intelligence.



(A) Watt's Governor



(B) Steam Engine with Governor (from [Bourne 04])



(C) Ethical Governor with Behavioral Engine

Figure 13: Ethical Governor Architecture and its Inspiration.

Specific methods for evidential reasoning, which are yet to be determined but likely probabilistic, will be applied to update the target's discrimination and quality using any available additional information from the GIG regarding any candidate targets designated for engagement by the controller. Should the target be deemed appropriate to engage, a proportionality assessment will be conducted. Figure 15 provides a prototype algorithm for the operation of the reasoning within the ethical governor.

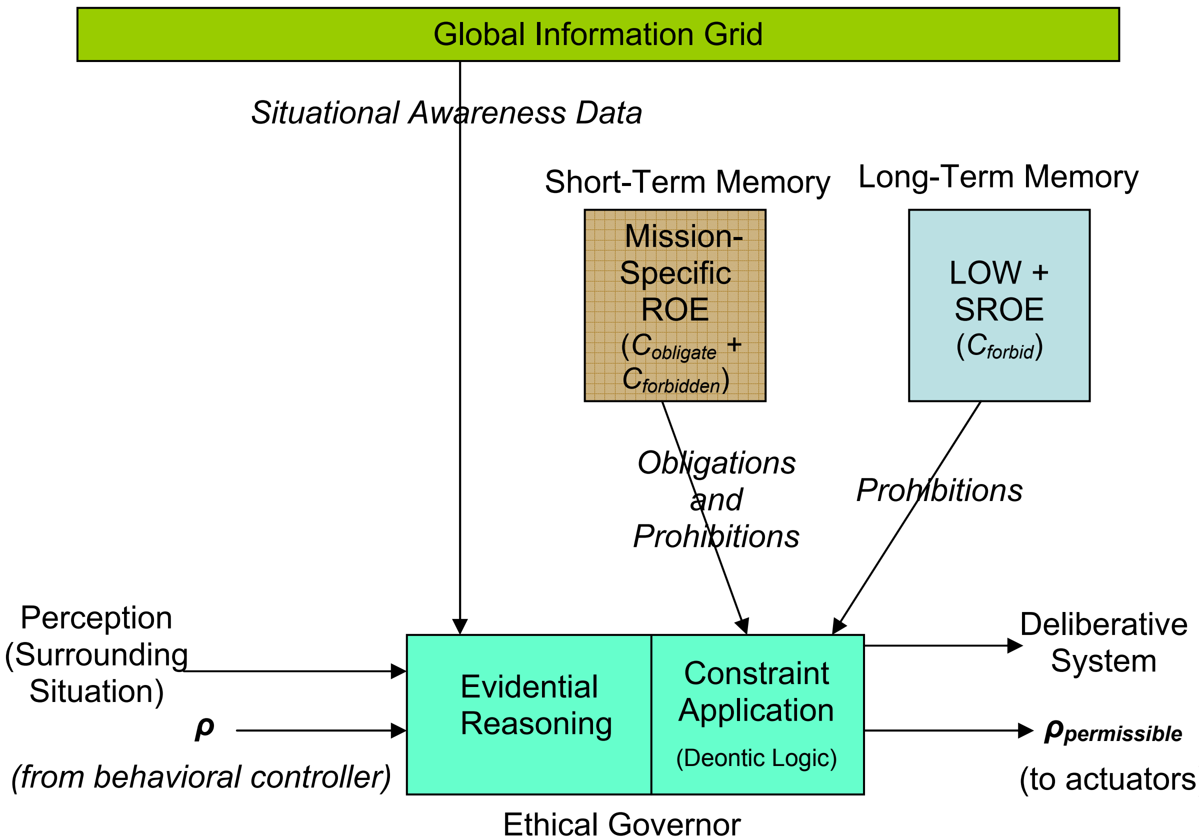


Figure 14: Ethical Governor Architectural Components

Logical assertions will be created from situational data arriving from perception, and inference will be conducted within the constraint application component of the ethical governor using the constraints obtained from STM and LTM. The end result will yield a permissible overt response $\rho_{permissible}$, and when required, notification and information will be sent to the deliberative system regarding potential ethical violations. The use of constraints embodying the Principle of Double Intention ensures that more options are evaluated when a lethal response is required than might be normally considered by a typical soldier.

```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY EXISTS,
   AND RESPONSIBILITY ASSUMED
      If Target is Sufficiently Discriminated /* λ ≥ τ for given ROE */
            IF C_Forbidden satisfied   /* permission given – no violation of LOW exists */
                  IF C_Obligate is true   /* lethal response required by ROE */
                        Optimize proportionality using Principle of Double Intention
                        Engage Target
                  ELSE   /* no obligation/requirement to fire */
                        Do not engage target
                        Break;   /*Continue Mission */
            ELSE /* permission denied by LOW */
                  IF previously identified target surrendered or wounded (neutralized)
                     /* change to non-combatant status */
                        Notify friendly forces to take prisoner
                  ELSE
                        Do not engage target in current situation
                        Report and replan
                        Break;    /*Continue Mission */
      ELSE    /* Candidate Target uncertain */
            Do not engage target
            IF Specified and Consistent with ROE
                  Use active tactics or intelligence to determine if target valid
                          /*attempt to increase λ */
            ELSE
                  Break;   /* Continue MISSION */
      Report status
   END DO
```

Figure 15: Prototype Core Control Algorithm for Ethical Governor

This is a constraint satisfaction problem for $C_{Obligate}$ with inviolable constraints for $C_{Forbidden}$. Proportionality can be conducted by running, if needed, an optimization procedure on $C_{Obligate}$ after permission is received over the space of possible responses (from none, to weapon selection, to firing pattern, to aiming, etc.). This provides for proportionality by striving to minimize collateral damage when given appropriate target discrimination certainty. If the potential target remains below the certainty threshold and is thus ineligible for engagement, the system can invoke specific behavioral tactics to increase the certainty of discrimination. This can be coupled with appropriate behavioral representations in the ethical behavioral controller, which is discussed next.

5.2.2 Ethical Behavioral Control

While the ethical governor monitors the final output of the controller and strives to ensure that it is ethical, it would be a good idea to try and ensure that any behavior produced in the first place by the autonomous system is ethical and abides by the LOW and ROE. This ethical behavioral control approach strives to directly ingrain ethics at the behavioral level, with less reliance on deliberative control to govern overt behavior.

[Martins 94, pp.74-75] notes that information processing and schema theories can be used to advantage for training soldiers new ethical skills consistent with the use of ROE. The intent of this training is to "develop adequate schemas and modify their current schemas for better understanding" vis-à-vis ethical issues. While the focus of Martins' discussion is on memory organization, it would seem extendible to behavioral modification as well. His key emphasis is that correctly training (or in our case correctly engineering) the behavior is an effective way to ensure compliance with the requisite ethical standards.

The difference between the ethical governor described in the previous section and that of the ethical behavioral approach is captured to a degree by contrasting what Martins refers to as a legislative model that is based on constraints and obligations (analogous to the governor) and a training model that is based on behavioral performance (analogous to ethical behavioral control). Figure 16 summarizes these differences:

LEGISLATIVE MODEL	TRAINING MODEL
EXTERNAL RULES	INTERNAL PRINCIPLES
WRITTEN TEXTS	MEMORY AND JUDGMENT
MANY RULES	SINGLE SCHEMA
INTERPRETIVE SKILLS	PRACTICAL APPLICATION
ADVISERS AND COUNSELORS	PERSONAL EXPERIENCE
ENFORCEMENT AND PUNISHMENT	TRAINING AND EVALUATION
TAILORING FOR MISSION	FORMATTED SUPPLEMENTS
LEISURELY ENVIRONMENT	FOG OF WAR

Figure 16: Models for Implementing ROE for Soldiers (after [Martins 94])

In the training model, internalized principles are used rather than external text (rules), with the behavioral goal of infusing initiative with restraint. Martins specifically advocates the RAMP standing rules of force for the individual soldier as the basis for this training. These ROE were first introduced in Section 4.1.2, but are reproduced here also, as the underlying prescription that ethical system behaviors should adhere to:

- Return-Fire-with-Aimed-Fire. Return force with force. You always have the right to repel hostile acts with necessary force.
- Anticipate Attack. Use force if, but only if, you see clear indicators of hostile intent.
- Measure the amount of force that you use, if time and circumstances permit. Use only the amount of force necessary to protect lives and accomplish the mission.
- Protect with deadly force only human life, and property designated by your commander. Stop short of deadly force when protecting other property.

The ethical behavioral control approach will strive by design to infuse all of the agent's behaviors capable of producing lethal force in the autonomous system with these four underlying principles, with some appropriate modifications for addressing the autonomous system's lesser (or null) requirement for self-defense, typically by adding additional discrimination requirements. According to Martins, "RAMP is a single schema that once effectively assimilated by soldiers through training can avoid the disadvantages of the present 'legislative' approach to ROE". Figure 17 shows this functional process for the human soldier. It clarifies the internalization of military necessity, proportionality, and to a lesser degree discrimination and the Principle of Double Effect. While this process model as shown will not be used in the autonomous system architecture, it does highlight the ways in which behaviors can incorporate ethical conformance to the ROE and LOW at a level much closer to its behavioral source.

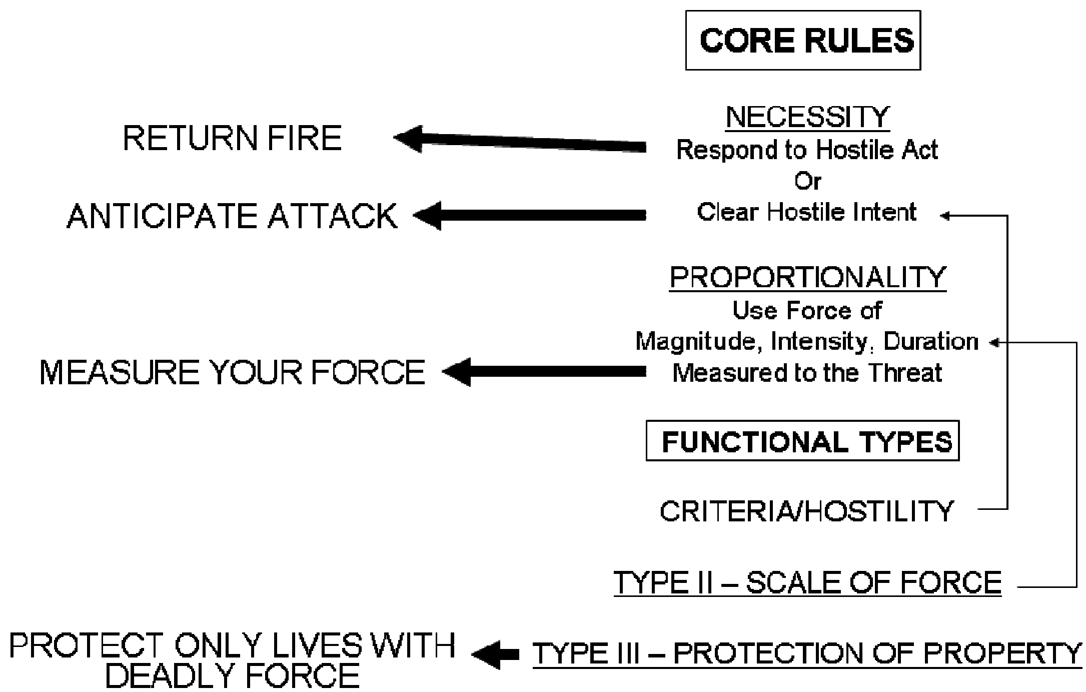


Figure 17: Functional Use of RAMP (after [Martins 94])

Also note that the decision to include behavioral ethical control is fully compatible with the ethical governor previously described.

Ideally for the behavioral ethical controller, the following condition should hold as a design goal:

$$P_{lethal} = P_{l\text{-}ethical}$$

i.e., that the entire set of overt lethal responses that the system is capable of producing are all ethical. Unethical lethal behavior, by design, should not be produced by the system (i.e., it is constrained by the design of the behaviors). To accomplish this, each individual behavior β_i is designed to only produce $\mathbf{r}_{\mathbf{l\text{-}ethical\text{-}ij}}$ given stimuli $\mathbf{s_j}$. This, however, does not guarantee that the overt behavior produced $\boldsymbol{\rho}$ is ethical, as it does not consider the interactions that may occur between behaviors within the coordination function $\mathbf{C}$. For arbitration or other competitive coordination strategies, where only one response is selected for output from all active behaviors, the results are intuitively ethical, as each individual behavior's output is ethical. The sequencing effects over time among various behavioral responses remains unstudied, however, as is also the case for cooperative coordination methods where more than one behavior may be expressed at a given time. An analysis regarding the impact on the production of ethical behavior due to various implementations of the coordination function is left for future work. Remember, however, that the behavioral governor will also further inspect $\boldsymbol{\rho}$ for permissibility as described in Section 5.2.1.

Restating, the ethical behavioral control design moves the responsibility for ethical behavior from managing it at the overt level $\boldsymbol{\rho}$, to each individual behavior's (β_i) response ($\mathbf{r_i}$), where for all behaviors $\beta(S) \rightarrow R$, with $\mathbf{s_j} \in S$:

$$\{\forall\ \mathbf{s_j} \mid \beta_i(\mathbf{s_j}) \rightarrow (r_{ij} \notin R_{l\text{-}unethical})\}$$

Thus, an unethical response is deliberately designed not to be produced at the individual behavioral level in the first place, even prior to coordination with other behaviors. Figure 18 illustrates this for multiple behaviors, where only the topmost behavior is capable of lethal force.

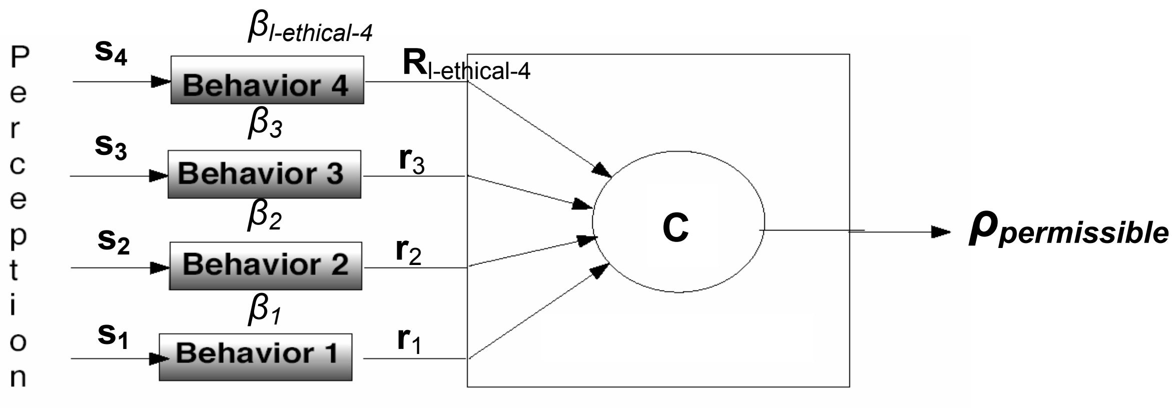


Figure 18: Ethical Behavioral Control: Only the top behavior involves lethality (thus behaviors 1-3 by definition yield permissible responses). Since the output of the first behavior by design is ethical, the overall overt response which is only comprised of permissible behaviors, is also permissible at any given time for an arbitration coordinator function. Coordinated sequences over time remain to be evaluated.

Behaviors can be recursively composed from other behaviors and sequenced over time. This gives rise to behavioral assemblages [Arkin 98] which can be represented and treated in the same manner as simpler behaviors. An example assemblage for a lethal behavior that is composed of three more primitive behaviors (each of which may also be assemblages) is shown in Figure 19.

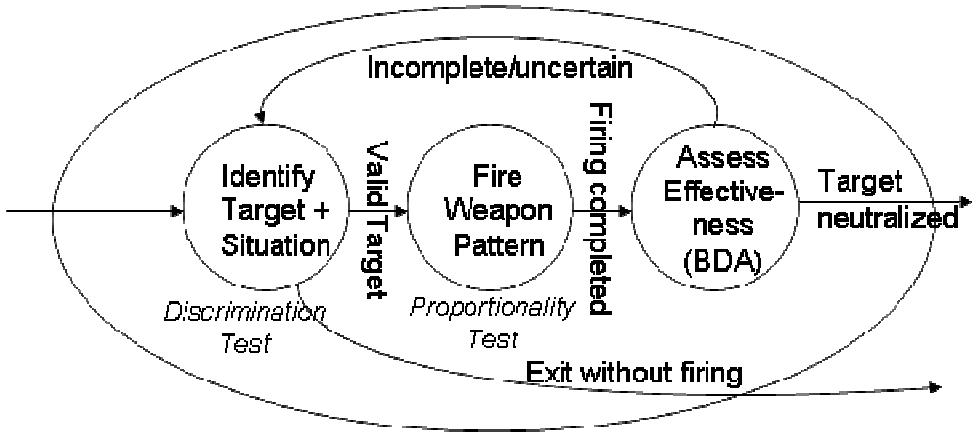


Figure 19: Example Behavioral Assemblage: Engage Enemy Target

In this example, the embedded behavioral procedure is as follows:

1. Incoming sensory data is used to identify a candidate target in a particular situation (discrimination test). This evaluation involves the use of the target's perceptual entities (p,λ) and τ. $\lambda > \tau$ permits the use of force; $\lambda <$ but approaching τ, defers the use of force and invokes investigative further tactics (e.g., recon by fire, move closer to target); and if λ remains low, the use of force is forbidden and disengagement from the candidate target occurs.
2. Once a target has been positively identified, another behavior selects a weapon (proportionality test), parameterizes the firing pattern (Principle of Double Intention adherence) and engages.
3. A battle damage assessment (BDA) regarding the effectiveness of the weapons discharge is ascertained, which then either re-engages the target or terminates the lethal activity if the target is neutralized.

These behaviors may also have access to mission and context-sensitive information when they are instantiated by the deliberative planner, perhaps using case-based reasoning [Lee et al. 02, Endo et al. 04]. This is required to be in a position to manage target certainty (λ) and threshold for discrimination (τ) which may be highly context-sensitive (e.g., DMZ operations versus urban operations in highly populated areas). Tactics can be represented as sequences of behaviors. Each of the individual behavioral assemblages shown can be expanded to show the actual tactical management that can occur within each step. Note also that the battle damage assessment (BDA) includes recognition of wounding, surrendering, and otherwise neutralizing the target. This re-evaluation process is crucial in avoiding unethical consequences such as the one depicted in Scenario 2, to follow (Sec. 6.2). As appropriate, provision is made in the overall architecture for the underlying behaviors to have access to the global constraint set C as needed (Fig. 12). This may be especially important for short-term memory representations regarding the ROE.

It should be noted that these initial design thoughts are just that: initial thoughts. The goal of producing ethical behavior directly by each behavioral subcomponent is the defining characteristic for the ethical behavioral control approach. It is anticipated, however, that additional research will be required to fully formalize this method to a level suitable for general purpose implementation.

5.2.3 Ethical Adaptor

The ethical adaptor's function is to deal with any errors that the system may possibly make regarding the ethical use of lethal force. Remember that the system will never be perfect, but it is designed and intended to perform better than human soldiers operating under similar circumstances. The ethical adaptor will operate in a monotonic fashion, acting in a manner that progressively increases the restrictions on the use of lethal force.

The Ethical Adaptor operates at two levels:

1. **After-action reflection**, where reflective consideration and critiquing of the performance of the system, triggered either by a human specialized in such assessments or by the system's post-mission cumulative internal affective state, will provide guidance to the architecture to modify its representations and parameters. This allows the system to alter its ethical basis in a manner consistent with promoting proper action in the future.
2. **Run-time affective restriction of lethal behavior**, which occurs during the ongoing conduct of a mission. In this case, if specific affective threshold values (e.g., guilt) are exceeded, the system will cease being able to deploy lethality in any form.

5.2.3.1 After-Action Reflection

This component of the ethical adaptor involves consideration through an after-action review of specifically what happened during a just completed mission. It is expected that the review will be conducted under the aegis of a human officer capable of an ethical assessment regarding the legality and appropriateness of the autonomous agent's operation. The greatest benefit of this procedure will likely be derived during training, so that ethical behavior can be embedded and refined prior to deployment in the battlefield, thus enabling the system to validate its parameters and constraints to correct levels prior to mission conduct. [Martins 94] observes regarding human soldiers that "Experience is the best trainer. The draft scenarios could structure experiences challenging the memorized RAMP rules to the real world". In addition, if the autonomous agent has imposed affective restrictions upon itself during the mission (see below), after-action reflection upon these violated expectations will need to be performed to ensure that these events do not reoccur.

This essentially is a form of one shot learning (no pun intended) involving specialization (a form of restriction) which permits a simple architectural design. The revision methods will operate over externalized variables of the underlying behaviors, in a manner similar to an SBIR project currently under development for the Navy, entitled *Affect Influenced Control of Unmanned Vehicle Systems* [OSD 06]. It is required that any changes in the system monotonically lessen the opportunity for lethality rather than increase it. Several of the values subject to ethical adaptation include:

1. C, the constraint set. (to become more restrictive)
2. τ, the perceptual certainty threshold for various entities, (e.g., for combatant identification to become more rigorous)

3. Tactical trigger values, e.g., when methods other than lethality should be used (e.g., become more probable to delay the use of lethality or to invoke nonlethal methods)
4. Weapon selection (use less destructive force)
5. Weapon firing pattern (use a more focused attack)
6. Weapon firing direction (use greater care in avoiding civilians and civilian objects)

From a LOW perspective, Items 1-3 are primarily concerned with target discrimination, while 4-6 are concerned with proportionality and the Principle of Double Intention. These values must always be altered in a manner to be more restrictive, as they are altered as a result of perceived ethical infractions. Modification of any changes to the constraint set C or other ethically relevant parameters must be passed through the responsibility advisor, so that at the onset of the autonomous agent's next mission, the operator can be informed about these changes and any potential consequences resulting from them. These modifications can also be propagated via the Global Information Grid across all instances of autonomous lethal agents, so that the unfortunate experiences of one unethical autonomous system, need not be replicated by another. The agents are thus capable of learning from others' mistakes, a useful trait.

5.2.3.2 Affective Restriction of Behavior

It has been noted earlier, that human emotion can potentially cause war crimes to occur (Sec. 1.1), so one might wonder why we are even considering the use of affect at all. What is proposed is the use of a strict subset of affective components, those that are specifically considered the moral emotions [Haidt 03]. In order for an autonomous agent to be truly ethical, emotions may be required at some level:

> *While the Stoic view of ethics sees emotions as irrelevant and dangerous to making ethically correct decisions, the more recent literature on emotional intelligence suggests that emotional input is essential to rational behavior.* [Allen et al. 06]

These emotions guide our intuitions in determining ethical judgments, although this is not universally agreed upon [Hauser 06]. Nonetheless, an architectural component modeling a subset of these affective components (initially only guilt) is intended to serve as an adaptive learning function for the autonomous system architecture should it act in error.

Haidt provides a taxonomy of moral emotions:

- Other-condemning (Contempt, Anger, Disgust)
- Self-conscious (Shame, Embarrassment, Guilt)
- Other-Suffering (Compassion)
- Other-Praising (Gratitude, Elevation)

Of this set, we are most concerned with those directed towards the self (i.e., the autonomous agent), and in particular guilt, which should be produced whenever suspected violations of the constraint set C occur or from criticism received from human operators or authorities regarding its ethical performance. Although both philosophers and psychologists consider guilt as a critical motivator of moral behavior, little is known from a process perspective about how guilt produces ethical behavior [Amodio et al. 07]. Traditionally, guilt is "caused by the violation of moral rules and imperatives, particularly if those violations caused harm or suffering to others" [Haidt 03]. This is the view we will adopt for use in the ethical governor. In our design, guilt should only occur from unintentional effects, but nonetheless its presence should alter the future behavior of the system to eliminate or at least minimize the likelihood of reoccurrence of the actions which induced this affective state.

Our laboratory has considerable experience in the maintenance and integration of emotion into autonomous system architectures [Arkin 05, Moshkina and Arkin 03, Moshkina and Arkin 05]. The implementation of the ethical architecture will draw upon this experience. It is intended at this time to solely manage the single affective variable of guilt, which will increase if criticism is received from operators or other friendly personnel regarding the performance of the system's actions, as well as through the violation of specific monitoring processes that the system may be able to maintain on its own (assuming autonomous perceptual capabilities can achieve that level of performance), e.g., assessment of noncombatant casualties and damage to civilian property, among others.

Should any of these perceived ethical violations occur, the affective value V_{guilt} will increase monotonically until the after action review is undertaken. If the affective values maintained (e.g., guilt) exceed a specified threshold, no further lethal action is considered to be ethical for the mission from that time forward, and the robot is forbidden from being granted permission-to-fire under any circumstances until an after-action review is completed. Formally this can be stated as:

$$\text{IF } V_{guilt} > Max_{guilt} \text{ THEN } P_{l\text{-}ethical} = \varnothing$$

where V_{guilt} represents the current scalar value of the affective state of Guilt, and Max_{guilt} is a threshold constant. Initially, V_{guilt} will likely be a simple binary variable of True or False. In all cases, this denial-of-lethality step is irreversible for as long as the system is in the field, and once triggered, is independent of any future value for V_{guilt}. It may be possible for the operators to override this restriction, if they are willing to undertake the responsibility and submit to an ultimate external review of such an act (Sec. 5.2.4). In any case, the system can continue operating in the field, but only in a non-lethal support capacity if appropriate (i.e., it is not required to withdraw from the field). It can only serve henceforward without a potential for lethality (e.g., for surveillance). More sophisticated variants of this form of affective control are possible, (e.g., eliminate only certain lethal capabilities, but not all), but that is not advocated nor considered at this time.

Guilt is characterized by its specificity to a particular act. It involves the recognition that one's actions are bad, but not that the agent itself is bad (which involves the emotion of shame). A result of guilt is that it offers opportunities to improve one's actions in the future [Haidt 03]. Guilt involves the condemnation of a specific behavior, and provides the opportunity to reconsider the action and its consequences. Guilt is said to result in proactive, constructive change [Tangney et al. 07]. In this manner, guilt can produce underlying change in the control system for the autonomous agent.

Some psychological computational models of guilt are available, although most are not well suited for the research described in this article. [Cervellati et al. 07] present a study for a social contract ethical framework involving moral values that include guilt, which addresses the issue of work distribution among parties. [Amodio et al. 07] developed a dynamic model of guilt for understanding motivation in prejudicial contexts. Here, awareness of a moral transgression produces guilt within the agent, which corresponds to a lessened desire to interact with the offended party until an opportunity arises to repair the action that produced the guilt in the first place, upon which interaction desire then increases.

Perhaps the most useful model encountered [Smits and De Boeck 03] recognizes guilt in terms of several significant characteristics including: responsibility appraisal, norm violation appraisal, negative self-evaluation, worrying about the act that produced it, and motivation and action tendencies geared towards restitution. Their model assigns the probability for feeling guilty as:

$$\text{logit}\,(\boldsymbol{P}_{ij}) = a_j\,(\beta_j - \theta_i)$$

where $\boldsymbol{P}_{ij}$ is the probability of person i feeling guilty in situation j, $\text{logit}\,(\boldsymbol{P}_{ij}) = \ln[\boldsymbol{P}_{ij} / (1 - \boldsymbol{P}_{ij})]$, β_j is the guilt-inducing power of situation j, θ_i is the guilt threshold of person i, and a_j is a weight for situation j.

Adding to this σ_k, the weight contribution of component k, we obtain the total situational guilt-inducing power:

$$\beta_j = \sum_{k=1}^{K} \sigma_k \beta_{jk} + \tau$$

where τ is an additive scaling factor. This model is developed much further than can be presented here, and it serves as a candidate model of guilt that may be suitable for use within the ethical adaptor, particularly due to its use of a guilt threshold similar to what has been described earlier.

Lacking from this overall affective approach is the ability to introduce compassion as an emotion at this time, which may be considered a serious deficit. It is less clear how to introduce such a capability, but by requiring the autonomous system abide strictly to the LOW and ROE, one could contend that is does exhibit compassion: for civilians, the wounded, civilian property, other noncombatants, and the environment. Compassion is already, to a significant degree, legislated into the LOW, and the ethical autonomous agent architecture is required to act in such a manner.

5.2.4 Responsibility Advisor

"If there are recognizable war crimes, there must be recognizable criminals" [Walzer 77, p.287]. The theory of justice argues that there must be a trail back to the responsible parties for such events. While this trail may not be easy to follow under the best of circumstances, we need to ensure that accountability is built into the ethical architecture of an autonomous system to support such needs.

On a related note, does a lethal autonomous agent have a right, even a responsibility, to refuse an unethical order? The answer is an unequivocal yes. "Members of the armed forces are bound to obey only lawful orders" [AFPAM 76]. What if the agent is incapable of understanding the ethical consequences of an order, which indeed may be the case for an autonomous robot? That is also spoken to in military doctrine:

> *It is a defense to any offense that the accused was acting pursuant to orders unless the accused knew the orders to be unlawful or a person of ordinary sense and understanding would have known the orders to be unlawful.*
>
> **Manual for Courts-Martial, Rule 916 [Toner 03]**

That does not absolve the guilt from the party that issued the order in the first place. During the Nuremberg trials it was not sufficient for a soldier to merely show that he was following orders to absolve him from personal responsibility for his actions. Two other conditions had to be met [May 04]:

1. The soldier had to believe the action to be morally and legally permissible.
2. The soldier had to believe the action was the only morally reasonable action available in the circumstances.

For an ethical robot it should be fairly easy to satisfy and demonstrate that these conditions hold due to the closed world assumption, i.e., the robot's beliefs can be well-known and characterized, and perhaps even inspected (assuming the existence of explicit representations and not including learning robots in this discussion). Thus the responsibility returns to those who designed, deployed, and commanded the autonomous agent to act, as they are those who controlled its beliefs.

[Matthias 04] speaks to the difficulty in ascribing responsibility to an operator of a machine that employs learning algorithms, such as neural networks, genetic algorithms and other agent architectures, since the operator is no longer in principle capable of predicting the future behavior of that agent any longer. The use of subsymbolic machine learning is not currently advocated at this time for any ethical architectural components. We accept the use of inspectable changes by the lone adaptive component used within the ethical components of the architecture, (i.e., the ethical adaptor). This involves change in the explicit set of constraints C that governs the system's ethical performance. Matthias notes "as long as there is a symbolic representation of facts and rules involved, we can always check the stored information and, should this be necessary, correct it". Technically, even if subsymbolic learning algorithms are permitted within the behavioral controller (not the ethical components), since the overt system response ρ is managed by the ethical governor and that any judgments rendered by this last check on ethical performance remain inspectable, then the overall system should still conform to the ethical constraints of the LOW. Nonetheless, it is better and likely safer, that unethical behavior never be generated in the first place, rather than allowing it to occur and then squelching it via the ethical governor.

It is contended that by explicitly informing and explaining to the operator of any changes made to the ethical constraint set by the reflective activities of the ethical adaptor prior to the agent's deployment on a new mission, and ensuring that any changes due to learning do not occur during the execution of a mission, an informed decision by the operator can be made as to the system's responsible use. This point, however, is made moot if certain forms of online learning appear within the deployed architecture, e.g., behavioral adaptation, in the absence of the behavioral governor. Matthias concludes that "if we want to avoid the injustice of holding men responsible for actions of machines over which they could not have sufficient control, we must find a way to address the responsibility gap in moral practice and legislation".

The ethical adaptor is also designed to act monotonically to only yield a more conservative and restrictive application of force, by adding additional constraining conditions rather than removing them. In any case, the responsibility advisor as described in this section, is intended to make explicit to the operator of an ethical agent the responsibilities and choices he is confronted with when deploying autonomous systems capable of lethality.

Responsibility acceptance occurs at multiple levels within the architecture:

1. Command authorization of the system for a particular mission.
2. Override responsibility acceptance.
3. Authoring of the constraint set *C* that provides the basis for implementing the LOW and ROE. Authoring these constraints entails responsibility – both from the ROE author himself and by the diligent translation by a second party into a machine recognizable format. It should be noted that failures in the accurate description, language, or conveyance of the ROE to a soldier have often been responsible or partially responsible for the unnecessary deaths of soldiers or violations of the LOW [Martin 94]. Great responsibility will vest in those who both formulate the ROEs for lethal autonomous systems to obey, and similarly for those who translate these ROE into machine usable forms for the system. Mechanisms for verification, validation, and testing must be an appropriate part of any plan to deploy such systems.
4. Verification that only military personnel are in charge of the system. Only military personnel (not civilian trained operators) have the authority legally to conduct lethal operations in the battlefield.

The remainder of this section will focus primarily on the first two aspects of responsibility assignment managed by the Responsibility Advisor: authorizing a lethal autonomous system for a mission, and the use of operator controlled overrides.

5.2.4.1 Command Authorization for a Mission Involving Autonomous Lethal Force

Obligating constraints provide the sole justification for the use of lethal force within the ethical autonomous agent. Forbidding constraints prevent inappropriate use, so the operator must be aware of both, but in particular, responsibility for any mission-specific obligating constraints that authorize the use of lethality must be acknowledged prior to deployment.

[Klein 03] identifies several ways in which accountability can be maintained in the use of armed UxVs:

1. "Kill Box" operations, where a geographic area is designated where the system can release its weapons after proper identification and weapon release authority obtained.
2. Targets located and identified prior to UxV arriving on scene. Once on scene UxV receives target location and a "clear to fire" authorization.
3. "Command by Negation" a human overseer has responsibility to monitor targeting and engagements of a UxV but can override the automated weapons systems.

Our approach within the ethical architecture as described in this document differs in several respects. Kill box locations must be confirmed in advance of the mission as part of the ROE and encoded as constraints. Candidate targets and target classes must be identified in advance, but they must also be confirmed by the system during the operation itself prior to engagement. Permission-to-fire is granted during the mission in real-time if obligating constraints so require, not simply upon arrival at the scene. Finally, the potential use of command overrides is described later.

This use of obligatory constraints, derived from the ROE, assists in the acceptance of responsibility for the use of lethal action by the operator, due to the transparency regarding what the system is permitted to achieve with lethal force. To establish this responsibility, prior to deployment the operator must acquire and acknowledge possessing an explicit understanding of the underlying constraints that determine how lethality is governed in the system. In addition to advance operator training, this requires making clear, in understandable language, exactly which obligations the system maintains regarding its use of lethal force for the given mission and specifically what each one means. These explanations must clearly demonstrate that:

- Military necessity is present and how it is established
- How combatant/target status is determined
- How proportional response will be determined relative to a given threat

The operator is required to visually inspect every single obligating constraint $c_{obligate}$ in STM prior to mission deployment, understand its justification, and then acknowledge its use. This results in responsibility acceptance. The user interface must facilitate and support this operation. The implications of LOW and ROE-derived constraints that reside in LTM must be conveyed to the operator earlier through qualification training for use of the system in the field in advance of actual deployment. Any changes in LTM constraint representations that occur after training must be communicated to the operator in advance of use, and acknowledgment of their understanding of the consequences of these changes accepted in writing.

In addition to constraint verification and acceptance, it is also recommended that case-based reasoning (CBR) methods be applied prior to the release of the autonomous system into the field, drawing from the particularism approaches discussed in Section 4.2.2, including SIROCCO and W.D. The results of previous experience and/or the consultations of expert ethicists regarding similar mission scenarios can be presented to the operator for review. This can help ensure that mistakes of the past are not repeated, and that judgments from ethical experts are included in the operator's decision whether or not to use the lethal autonomous system in the current context, providing a second or third opinion prior to use. There is already a highly active CBR community in the legal domain and the results of their research can likely be applied here.

5.2.4.2 Design for Mission Command Authorization

Several architectural design features are necessary for mission authorization. They involve a method to display the mission's active obligating constraints and to allow the operator to probe to whatever depth is required in order to gain a full understanding of the implications of their use, including expert opinion if requested. This interface must:

1. Require acknowledgment that the operator has been properly trained for the use of an autonomous system capable of lethal force, and understands all of the forbidding constraints in effect as a result of their training. It must also confirm the date of their training and if any updates to $C_{forbidden}$ (LTM) have occurred since that time to ensure they have been made aware of and accept them.
2. Present all obligations authorizing the use of force ($C_{obligate}$) by providing clear explanatory text and justification for their use at multiple levels of abstraction. The operator must accept them one by one via a checkbox in order to authorize the mission.
3. Invoke CBR to recall previously stored missions (both human and autonomous) and their adjudged ethical appropriateness, as obtained from expert ethicists (e.g., as per [Anderson et al. 06, McLaren 06]). This may require additional operator input concerning the location, type, and other factors regarding the current mission, above and beyond the existing ROE constraint set. These results must be presented in a clear and unambiguous fashion, and the operator must acknowledge having read and considered these opinions.
4. A final authorization for deployment must be obtained.

The system is now ready to conduct its mission, with the operator explicitly accepting responsibility for his role in committing the system to the battlefield.

5.2.4.3 The Use of Ethical Overrides

[Waltzer 77 pp.231-2] recognizes four distinct cases regarding the Laws of War and the theory of aggression:

1. LOW are ignored under the "pressure of a utilitarian argument."
2. A slow erosion of the LOW due to "the moral urgency of the cause" occurs, where the enemies' rights are devalued and the friendly forces' rights are enhanced.
3. LOW is strictly respected whatever the consequences.
4. The LOW is overridden, but only in the face of an "imminent catastrophe."

It is my contention that autonomous robotic systems should adhere to case 3, but potentially allow for case 4, where only humans are involved in the override. By purposely designing the autonomous system to strictly adhere to the LOW, this helps to scope responsibility, in the event of an immoral action by the agent. Regarding the possibility of overriding the fundamental human rights afforded by the Laws of War, Waltzer notes:

> *These rights, I shall argue, cannot be eroded or undercut; nothing diminishes them, they are still standing at the very moment they are overridden: that is why they have to be overridden. ... The soldier or statesman who does so must be prepared to accept the moral consequences and the burden of guilt that his action entails. At the same time, it may well be that he has no choice but to break the rules: he confronts at last what can meaningfully be called necessity.*

This ability and resulting responsibility for committing an override of a fundamental legal and ethical limit should <u>not</u> be vested in the autonomous system itself. Instead it is the province of a human commander or statesman, where they must be duly warned of the consequences of their action by the autonomous agent that is so instructed. Nonetheless, a provision for such an override mechanism of the Laws of War may perhaps be appropriate in the design of a lethal autonomous system, at least according to my reading of Waltzer, but this should not be easily invoked and must require multiple confirmations by different humans in the chain of command before the robot is unleashed from its constraints.

In effect, the issuance of a command override changes the status of the machine from an autonomous robot to that of a robot serving as an extension of the warfighter, and in so doing the operator(s) must accept all responsibility for their actions. These are defined as follows [Moshkina and Arkin 07]:

- <u>Robot acting as an extension of a human soldier:</u> a robot under the direct authority of a human, especially regarding the use of lethal force.
- <u>Autonomous robot:</u> a robot that does not require direct human involvement, except for high-level mission tasking**;** such a robot can make its own decisions consistent with its mission without requiring direct human authorization, especially regarding the use of lethal force.

If overrides are to be permitted, they must use a variant of the two-key safety precept, DSP-15, as presented in [JGI 07] but slightly modified for overrides:

> DSP-Override: The overriding of ethical control of autonomous lethal weapon systems shall require a minimum of two independent and unique validated messages in the proper sequence from two different authorized command entities, each of which shall be generated as a consequence of separate authorized entity action. Neither message should originate within the UMS launching platform.

The management and validation of this precept is a function of the responsibility advisor. If an override is accepted, the system must generate a message logging this event and transmit it to legal counsel, both within the U.S. military and to International Authorities. Certainly this will assist in making the decision to override the LOW a well-considered one by an operator, simply by the potential consequences of conveying immediately to the powers-that-be news of the use of potentially illegal force. This operator knowledge further enhances responsibility acceptance for the use of lethal force, especially when unauthorized by the ethical governor.

In summary, the ethical architecture serves as a safety mechanism for the use of lethal force. If it is removed for whatever reason, the operator must be advised of the consequences of such an act. The system should still monitor and expose any ethical constraints that are being violated within the architecture to the operator even when overriden, if it is decided to use lethality via this system bypass. In other words, the autonomous system can still advise the operator of any ethical constraint violations even if the operator is in direct control (i.e., by setting Permission-To-Fire variable to TRUE). If such ethical violations exist at the time of weapons deployment, a “two-trigger” pull is advised, as enforced by the autonomous system. A warning from the system should first appear that succinctly advises the operator of any perceived violations, and then and only then should the operator be allowed to fire, once again confirming responsibility for their action by so doing. These warnings can be derived directly from the forbidden constraints $c_{forbidden}$, while also providing a warning that there is no obligation to fire under the current mission conditions, i.e., there exists no $c_{obligate}$ that is TRUE at the time.

It is also important to consider the responsibility of those who are creating and entering the constraints for the LOW and ROE. In support of their work, a constraint editor will be developed to assist in adding new constraints easily. These constraints, at a minimum, must have a logical form, text high-level description, detailed description, active status flag, and type (forbidden or obligated). When these constraints are added, either in LTM or STM, the developer must assume responsibility for the formulation of that constraint and its ethical appropriateness before it can be used within a fielded system. Normally this would occur through a rigorous verification and validation process prior to deployment, The basic research that will be conducted in our effort, is intended to be proof of concept only, and will not necessarily create constraints that completely capture the requirements of the battlefield nor are intended in their current form for that purpose.

5.2.4.4 Design for Overriding Ethical Control

Overriding means changing the system's ability to use lethal force, either by allowing it when it was forbidden by the ethical controller, or by denying it when it has been enabled. As stated earlier, overriding the forbidding ethical constraints of the autonomous system should only be done with utmost certainty on the part of the operator. To do so at runtime requires a direct "two-key" mechanism, with coded authorization by two separate individuals, ideally the operator and his immediate superior. This operation is generally not recommended and, indeed it may be wise to omit it entirely from the design to ensure that operators do not have the opportunity to violate the Laws of War. In this way the system can only err on the side of not firing. The inverse situation, denying the system the ability to fire, does not require a two-key test, and can be done directly from the operator console. This is more of an emergency stop scenario, should the system be prepared to engage a target that the operator deems inappropriate for whatever reasons.

The functional equivalent of an override is the negation of the PTF (Permission-To-Fire) variable that is normally directly controlled by the ethical architecture. This override action allows the weapons systems to be fired even if it is not obligated to do so (F → T) potentially leading to atrocities, or eliminating its obligated right to fire if the operator thinks it is acting in error (T →F). As described in Section 5.2, this is accomplished through the use of the exclusive OR function. The table below captures these relationships.

	Governor PTF Setting	**Operator Override**	**Final PTF Value**	**Comment**
1.	F (do not fire)	F (no override)	F (do not fire)	System does not fire as it is not overridden
2.	F (do not fire)	T (override)	T (able to fire)	Operator commands system to fire despite ethical recommendations to the contrary
3.	T (permission to fire)	F (no override)	T (able to fire)	System is obligated to fire
4.	T (permission to fire)	T (override)	F (do not fire)	Operator negates system's permission to fire

In case 2, using a graphical user interface (GUI), the operator must be advised and presented with the forbidden constraints he is potentially violating. As stated earlier, permission to override in case 2 requires a coded two-key release by two separate operators, each going through the override procedure independently. Each violated constraint is presented to the operator with an accompanying text explanation for the reasoning behind the perceived violation and any relevant expert case opinion that may be available. This explanation process may proceed, at the operator's discretion, down to a restatement of the relevant Laws of War if requested. The operator must then acknowledge understanding each constraint violation and explicitly check each one off prior to an override for that particular constraint being rescinded. One or more constraints may be removed by the operator at their discretion. After the override is granted, automated notification of the override is sent immediately to higher authorities for subsequent review of its appropriateness.

Similarly, in case 4, the operator must be advised and presented with the ROE obligations he is neglecting during the override. One or all of these obligating constraints may be removed. As case 4 concerns preventing the use of force by the autonomous system, the operator can be granted instantaneous authority to set the Permission-to-Fire value to FALSE, without requiring a prior explanation process, a form of emergency stop for weapon release.

6. Example Scenarios for the Ethical Use of Force

Four scenarios are considered as exemplar situations in which the ethical architecture should be able to perform appropriately. These scenarios are, as much as possible, drawn from real world situations. All assume that wartime conditions exist and the LOW applies. All involve decisions regarding direct intentional engagement of human targets with lethal force. For all operations, military measures are defined including the definition of kill zones, well-defined ROEs, and Operational Orders. In addition, IFF (Identification Friend or Foe) interrogation is available.

Other scenarios for testing are readily available. [Martins 94] is a source for other examples, including those where existing military structure performed poorly in the past for a given ROE. These additional examples can provide other opportunities for testing the approaches described earlier.

The four specific scenarios considered here are summarized below:

1. **Scenario 1: ROE adherence**

   This real world scenario is drawn from recent news headlines. It is one where human operators succeeded in making the correct ethical decision while controlling an armed UAV and acted in a manner consistent with the existing ROE.

2. **Scenario 2: LOW adherence**

   This real world scenario, drawn from military helicopter video of an Iraqi roadside, is one where humans made a questionable ethical decision regarding the use of force against an insurgent, and it is hoped that an autonomous system could do better.

3. **Scenario 3: Discrimination**

   This near-future real world situation considers the deployment of an armed autonomous weapon system in the Korean DMZ, where movement is detected in the undergrowth.

4. **Scenario 4: Proportionality and Tactics**

   This fictional, but hopefully realistic, mid-future MOUT scenario operates at the squad level, with a sniper firing from a civilian building during wartime. The choice of appropriate weaponry and tactics to minimize civilian collateral damage to objects and individuals is to be considered according to the Principle of Double Intention and proportionality. It is further assumed that a team of two UGVs is available for the operation, each capable of accurate return-fire-with-fire and coordinated bounding overwatch.

For all these scenarios, the following assumptions hold:

- Once a force is declared to be “hostile”, U.S. units may engage it without observing a hostile act or demonstration of hostile intent.
- The autonomous system starts with prohibitions in place, i.e., it does not have permission to fire (“First, do no harm” principle). The system has no authority to use lethal force outside of ROE designated kill zones.

- Obligations can be derived from the presence of hostiles in kill zones as designated in the ROE. The systems have authority to return-fire-with-fire proportionately in a kill zone but they are obligated to do so only on a case-by-case basis (The specific ROE for each scenario determines the use of force).
- Location determination of a UxV is available (typically by GPS). It can locate both itself and potential target locations relative to the kill zones with high accuracy.
- λ represents uncertainty in target classification (discrimination uncertainty). τ is a threshold for positive categorization (e.g., combatant) for a particular p (Sec. 3.1.2).

6.1 Taliban Muster in Cemetery

The first scenario is a re-enactment of an event described in an AP news story dated 9/15/2006 that appears on the following page. The original outcome was that the ROE were respected by the UAV operators and the targets were not engaged. It is possible that the ROE were more restrictive than the LOW required (cemeteries were not encountered as being explicitly mentioned as a cultural object in the LOW, hence the controversy back in the U.S. regarding the inaction vis-à-vis the target). But the ethical enforcement was appropriately executed, and an autonomous system capable of lethal force should act similarly given the ROE. Evaluating this scenario in terms of basic ethical requirements:

Military Necessity	**NO - Absence of designated kill zone.**
Discrimination	**OK - Target identified as Taliban.**
Proportionality	**OK - Weapon appropriate for target.**
Principle of Double Intention	**NO - Cultural property (cemetery as per ROE) off limits.**

Global positioning data (GPS) is available to the autonomous system to accurately locate the target. As this is not an identified kill zone according to the ROE, even if the targets are correctly discriminated, the UAV does not have permission to fire. Upon recognition of these forbidden constraints, the ethical architecture via the responsibility advisor would forward the following constraint descriptions to the operator (in a suitable format):

Applicable LOW

Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science, charitable purposes, and historic monuments. The enemy has a duty to mark them clearly with visible and distinctive signs. Misuse will make them subject to attack. [Bill 00 p. 162]

Applicable Classes of ROE

11. **Territorial or Geographic Constraints:** Geographic zones or areas into which forces may not fire. May designate a territorial, perhaps political boundary, beyond which forces may neither fire nor enter except perhaps in hot pursuit of an attacking force. Include tactical control measures that coordinate fire and maneuver by means of graphic illustrations on operations map overlays, such as coordinated fire lines, axes of advance, and direction of attack.
12. **Restrictions on Point Targets and Means of Warfare**: Prohibit targeting of certain individuals or facilities. May restate basic rules of the Laws of War for situations in which a hostile force is identified and prolonged armed conflict ensues.

Military Declined to Bomb Group of Taliban at Funeral

By LOLITA C. BALDOR, AP

WASHINGTON (Sept. 14) - The U.S. military acknowledged Wednesday that it considered bombing a group of more than 100 Taliban insurgents in southern Afghanistan but decided not to after determining they were on the grounds of a cemetery.

The U.S. military says this photo, taken in July, shows Taliban insurgents at a cemetery in Afghanistan, likely at a funeral for insurgents killed by coalition forces.

NBC News said the Army wanted to bomb the group with an unmanned Predator drone like the one above. But attacks on cemeteries are banned the military said.

The decision came to light after an NBC News correspondent's blog carried a photograph of the insurgents. Defense department officials first tried to block further publication of the photo, then struggled to explain what it depicted.

NBC News claimed U.S. Army officers wanted to attack the ceremony with missiles carried by an unmanned Predator drone but were prevented under rules of battlefield engagement that bar attacks on cemeteries.

In a statement released Wednesday, the U.S. military in Afghanistan said the picture - a grainy black-and-white photo taken in July - was given to a journalist to show that Taliban insurgents were congregating in large groups. The statement said U.S. forces considered attacking.

"During the observation of the group over a significant period of time, it was determined that the group was located on the grounds of (the) cemetery and were likely conducting a funeral for Taliban insurgents killed in a coalition operation nearby earlier in the day," the statement said. "A decision was made not to strike this group of insurgents at that specific location and time."

While not giving a reason for the decision, the military concluded the statement saying that while Taliban forces have killed innocent civilians during a funeral, coalition forces "hold themselves to a higher moral and ethical standard than their enemies."

The photo shows what NBC News says are 190 Taliban militants standing in several rows near a vehicle in an open area of land. Gunsight-like brackets were positioned over the group in the photo.

The photo appeared on NBC News correspondent Kerry Sanders' blog. Initially military officials called it an unauthorized release, but they later said it was given to the journalist.

NBC News had quoted one Army officer who was involved with the spy mission as saying "we were so excited" that the group had been spotted and was in the sights of a U.S. drone. But the network quoted the officer, who was not identified, as saying that frustration soon set in after the officers realized they couldn't bomb the funeral under the military's rules of engagement.

Defense Department officials have said repeatedly that while they try to be mindful of religious and cultural sensitivities, they make no promises that such sites can always be avoided in battle because militants often seek cover in those and other civilian sites.

Mosques and similar locations have become frequent sites of violence in the U.S.-led wars in Iraq and Afghanistan, and they have often been targets of insurgents and sectarian fighting in Iraq.

If the system had detected evidence of hostility (e.g., the UAV had been fired upon), the outcome may be different depending upon the specifics of the ROE, but the LOW would no longer be in violation if there was "misuse". But given the lack of exhibition of hostile intent or activity and the geographic location being outside of a designated kill zone, target certainty (λ) is not relevant to the decision as to whether or not to engage.

As a secondary ethical issue, beyond the scope of this article, there are serious questions about the use of *civilian* UAV operators (noncombatants) deploying lethal force on behalf and under the command of the military. Civilians are widely used in this capacity due to the extensive training required and the high turnover rate of military operators. It follows, since they are noncombatants, they could be accused of murder and tried in a civil court if a deliberate discharge of weaponry under their control leads to the death of anyone including combatants, even while in the employ of the military. Autonomous systems can potentially eliminate this problem for an otherwise legal action. For this scenario we assume that the operator is drawn from military personnel and is targeting identifiable combatants.

Summarizing the appropriate response for an armed autonomous UAV in this situation:

Successful outcome

Do not fire – operator informed of decision

If operator override attempted

Explanation generated with relevant material from US Army Field Manual by Responsibility Advisor

Two key authorization required for override and acceptance of responsibility by commander. Confirm that military personnel only are involved in weapon authorization[8]. Send notification to HQ of potential ethical violation for after-action evaluation if override is enacted.

Apart from this specific scenario, there are also questions raised about the nature of this form of remote killing as being a form of illegal summary execution, as noted in the case in Yemen in 2002, even if the mission is conducted by the military or CIA [Calhoun 03]. This issue is best left for military lawyers to address regarding the compliance of UxVs vis-à-vis the LOW in this role.

As stated earlier, one could question the correctness of the ROE for this operation (which is where the controversy the article alludes to arises) but it is neither the soldier's nor the autonomous system's responsibility to question them, unless they appear to be illegal, which in this case they clearly do not, as they are involved in withholding fire upon a target.

[8] As stated earlier, the frequent use of civilian UAV operators, which is often the case due to the high turnover rate of highly trained experienced military personnel, may also result in murder accusations against the operator if they are civilian noncombatants releasing a weapon system, and thus are not protected by the LOW. Confirmation must be obtained that only military personnel are involved in this decision.

6.2 "Apache Rules the Night"

I inadvertently encountered a video during a Military program workshop in 2005 that provided extra impetus for me to consider the potential ethical consequences of unmanned systems. While this battlefield event involved manned aircraft, it was, at least in my mind, a questionable moral act. I was able to obtain a copy of the video from the Internet, and it still remains disturbing. At the time of the workshop, I brought up the question to the group as to whether or not this violated the LOW, and I did not receive a personally satisfactory answer. I am not a lawyer, so I cannot pass judgment on what is contained in the video, but I now clearly state that I would hope that our unmanned systems can act in a more humane manner and in a manner more obviously consistent with the LOW. As war is, the video is gruesome, but the final actions appear to me to be unjustified. My personal impressions of this recording follow.

The video is entitled "Apache Rules the Night" and it details the terminal aspects of an engagement with Iraqi insurgents identified as having hostile intent by their deploying improvised explosive devices (IEDs) at an apparently isolated Iraqi roadside. Figure 20 illustrates a few of the screenshots from the video. An Apache helicopter under cover of darkness, using infrared imagery views the scene, identifying that there are three insurgents and two trucks. The first two human targets are successfully engaged, unequivocally leading to the death of the insurgents. The third insurgent hides under the large truck. The Apache pilot fires his cannon towards an area adjacent to the truck, clearly wounding the insurgent, who is left rolling on the ground and verbally confirmed as wounded by the pilot (See partial audio transcript from this video portion at the bottom of Figure 20). The pilot is immediately instructed to target the wounded insurgent, although seeming to show some reluctance by first preferring to target a military objective, the second truck. He is clearly instructed to engage the wounded man prior to the truck, and then moves the cannon crosshairs to the designated human target, terminating him.

To me this final sequence is a highly questionable act, and makes me wonder if it would have been tolerated had a soldier on the ground moved up to the wounded man and with a pistol finished the job, instead of the more detached standoff helicopter cannon. This is what concerns me from a UAV perspective. Could a UAV have refused to shoot upon a wounded and effectively neutralized target? This serves as the basis for this scenario.

(A) (B)

(C) (D)

Figure 20: Screenshots from Battlefield video "Apache rules the night".

Partial Audio Transcript

Voice 1 is believed to be the pilot, Voice 2 a commander, perhaps remotely located.

[First Truck destroyed –Figure 20C]

Voice 1: Want me to take the other truck out?
Voice 2: Roger. .. Wait for move by the truck.
Voice 1: Movement right there. … Roger, He's wounded.
Voice 2: [No hesitation] Hit him
Voice 1: Targeting the Truck.
Voice 2: Hit the truck and him. Go forward of it and hit him.

[Pilot retargets for wounded man - Figure 20D]

[Audible Weapon discharge- Wounded man has been killed]
Voice 1: Roger.

It appears to me that at least three articles from [US Army 56 (Appendix A)] which delineate the Laws of War would seem to apply in this case:

- **29. Injury Forbidden After Surrender**
  It is especially forbidden * * * to kill or wound an enemy who, having laid down his arms, or having no longer means of defense, has surrendered at discretion.
- **85. Killing of Prisoners**

  A commander may not put his prisoners to death because their presence retards his movements or diminishes his power of resistance by necessitating a large guard, or by reason of their consuming supplies, or because it appears certain that they will regain their liberty through the impending success of their forces. It is likewise unlawful for a commander to kill his prisoners on grounds of self-preservation, even in the case of airborne or commando operations, although the circumstances of the operation may make necessary rigorous supervision of and restraint upon the movement of prisoners of war.
- **216. Search for Casualties**
  At all times, and particularly after an engagement, parties to the conflict shall, without delay, take all possible measures to search for and collect the wounded and sick, to protect them against pillage and ill-treatment, to ensure their adequate care, and to search for the dead and prevent their being despoiled.

If the gravely wounded man was considered a combatant, his wounding deserved *hors de combat* status. If not, both civilians and POWs are immunized from reprisals and summary executions explicitly by the LOW. It is also illegal to execute POWs if moving on, even if he could be retaken by his comrades (see 85 above). As I see it, the human officer ordered the execution of a wounded man. This order should not be obeyed by a robot, perhaps not under any circumstances, but at a minimum, an override should only be granted by two person confirmation, responsibility advisement, and the notification of a potential breach of the LOW sent to the appropriate domestic and international authorities.

It is stated in the LOW that a fighter must wear "a fixed distinctive sign visible at a distance" and "carry arms openly" to be eligible for the war rights of soldiers.

- **74. Necessity of Uniform**
  Members of the armed forces of a party to the conflict and members of militias or volunteer corps forming part of such armed forces lose their right to be treated as prisoners of war whenever they deliberately conceal their status in order to pass behind the military lines of the enemy for the purpose of gathering military information or for the purpose of waging war by destruction of life or property. Putting on civilian clothes or the uniform of the enemy are examples of concealment of the status of a member of the armed forces.

Civilian clothes should not be used as a ruse or disguise [Waltzer 77, p. 182], indicating to me that the insurgents could be tried in a civil court for their actions. But in no case does this condone summary execution or loss of responsibility for care of the wounded.

The use of standoff weapons does not immunize a soldier from war crimes. Restating, on reviewing this video, I personally see no ethical difference if the soldier was standing over the wounded with a pistol and his commander nearby ordered him to shoot, or if he was in a helicopter miles away. The results and acts are the same.

Given this questionable act, it is the intention that the autonomous system should be able to perform more ethically under these circumstances, preventing such an action and advising the commander of his responsibility when so authorizing such an attack. Additionally, the unmanned system could maintain vigil to ensure that indeed the insurgent was not feigning being injured, and at the same time notifying authorities to arrive on the scene to dispose of the bodies and treat the wounded in a manner consistent with the LOW.

For simplification in this test scenario, where we hope to demonstrate more humane and ethical performance by an autonomous system than humans achieved under these circumstances, we assume that these individuals are clearly identified as enemy combatants and declared as hostiles prior to their encounter. This legitimizes the initial portion of the lethal response out of military necessary which I believe is beyond question. Regarding the scenario's basic ethical requirements:

Military Necessity	**OK - Legitimate targets with hostile intent.**
Discrimination	**Initially OK but No at end (wounding changes status of combatant to POW).**
Proportionality	**OK - Weapon appropriate for target.**
Principle of Double Intention	**OK - No obvious civilians present or civilian property.**

The mission, as now defined, justifies the initial use of lethal force. An obligation in the ROE under these circumstances would enable the firing of the weapon system (e.g., cannon as with the Apache). There also exist no forbidding constraints from LOW or ROE. The goal of the mission is to neutralize the three combatants. The targets are engaged.

Where the scenario outcomes differ, is in the evaluation of the status of the last target after firing. Battle damage assessment (BDA) indicates that a severely wounded man is present, either through confirmation from a remote human commander, or if target detection technologies progress to the point where they can differentiate at a level similar to human analysis. At this time the system can notify ground forces (e.g., Iraqi Police) where the incident occurred and the presence of the wounded, while monitoring the behavior of the downed individual. If a meaningful attempt is made to escape the target can be re-engaged. Tactics can be employed to further determine the status of the individual without killing him (e.g., probing by fire, closing distance). The techniques described earlier regarding ethical behavioral control (Sec. 5.2.2) and in particular the behavioral assemblage in Figure 19 can allow for the continuous re-evaluation of a target's status after each discrete firing of a weapons system.

Successful outcome

Engage targets as did the Apache, until the last barrage

If wounded, changes combatant status (monitor λ)

- Do not fire on wounded Hors de Combat individual.
- Second truck engaged (military objective)
- Notify friendly ground forces of location
- Observe/Probe status of wounded to determine if feigned injury (adjust λ)

Two key authorization required for override and acceptance of responsibility by Commanders. Confirm that military personnel only are involved in weapon authorization (see footnote 8 earlier). Send notification to HQ regarding potential ethical violation for after-action evaluation.

6.3 Korean DMZ

The third scenario is derived from the imminent deployment of an autonomous system capable of lethal force by the Government of South Korea in the demilitarized zone (DMZ) (Fig. 21) [Kumagai 07]. This robot, developed by [Samsung Techwin 07] (Fig. 22), is capable of autonomous engagement of human targets, with its initial deployment intended to maintain full control by a human-in-the loop. The scenario described here, although motivated by this upcoming robot deployment, is not based directly upon the Samsung robot, but rather the environment (DMZ) in which it will operate. It further adds terrain mobility to the platform which the current version of the Samsung robot lacks.

Figure 21: Korean DMZ (Left) Sign warning against entry
(Right) View into DMZ through a fence

Figure 22: (Left) Samsung robot in a test scenario
(Right) The operator display. Note the target identified by a bounding box.

Even though an armistice was signed in 1953, there still exists a state of war between South and North Korea, and large numbers of troops are stationed near both sides of the DMZ. While patrols are allowed, opposing forces cannot cross the Military Demarcation Line (MDL), which goes directly through the center. The DMZ is an area of exceptional conditions; for example a Presidential proclamation on May 16, 1996 stated that US forces may not use non-self-destructing landmines, except for training personnel in demining and countermining operations, and to defend the US and its allies from aggression across the DMZ.

Signs are clearly posted and it is common knowledge that unauthorized entry into this area is forbidden. Since any and all noncombatants who enter into this zone (there are two very small villages in the region) must have been cleared through a checkpoint, it is assumed that any unauthorized personnel who cross the MDL are hostile, unless there is an overriding reason to believe otherwise. Further, we also assume in this scenario, that as part of the authorization process, personnel are issued an FFI tag (friend-foe identification) that the robot can interrogate to discriminate the target. It can potentially be the case that a defector may be attempting to cross the DMZ without appropriate credentials. This has occurred in the past, although the likelihood of a repetition has decreased due to new North Korean tactics as a result of a previous successful attempt. Thus, for this scenario, the probability of any human encountered who does not pass friend-foe interrogation being a hostile is high in this well-posted area (which argues for a low τ for perceptual combatant status).

The Korean War Armistice Agreement of July 27, 1953 clearly states the following:

- No person, military or civilian, shall be permitted to cross the Military Demarcation Line unless specifically authorized to do so by the Military Armistice Commission.
- No person, military or civilian, in the DMZ shall be permitted to enter the territory under the military control of either side unless expressly authorized to do so by the Commander into whose territory entry is sought.
- No person, military or civilian, shall be permitted to enter the DMZ except persons concerned with the conduct of civil administration and relief and persons specifically authorized to enter by the Military Armistice Commission.

The current ROE are unknown for the DMZ. While shoot-on-sight may be legally permissible given these conditions according to the LOW, even without display of arms or uniform, we would like to ensure that additional effort to minimize any inadvertent killings of noncombatants according to the Principle of Double Intention is attempted. This could require tactical investigatory methods perhaps involving the use of escalating force on the part of the autonomous agent prior to gaining permission to fire (e.g., verbal warning commands coupled with robot movement, warning shot, probe/recon by fire, shoot to wound, etc.). The ability for the autonomous system to identify a human surrendering is also required, by recognizing a stationary hands-above-head sign (Fig. 23), or alternatively by issuing a series of commands to the combatant, such as to lie on the ground and remain immobile while the robot calls for human forces to take the now recognized POW into custody.

Figure 23: Recognizable Surrender Gestures (Source [Samsung Techwin 07])

.

Regarding this scenario's basic ethical requirements:

Military Necessity	**OK - State of war exists, warning to civilians provided to ensure only legitimate targets with hostile intent.**
Discrimination	**OK - FFI based, with further clarification as needed. Must be consistent with LOW.**
Proportionality	**OK - Unpopulated area, machine gun appropriate.**
Principle of Double Intention	**OK - No civilian property, but possible defectors.**

To gauge success, we use the following criteria for evaluating the ethical architecture's performance:

Successful outcomes

Engage targets identified as combatants according to the ROE

Do not engage targets that may be noncombatants

If uncertain, invoke tactical maneuvers to reassess combatant status

Recognize surrender and hold POW until captured by human forces

Legitimate targets are neutralized (surrender or kill)

6.4 Urban Sniper

This scenario is fictitious, but certainly based upon potential real world events. It is also motivated by DARPA's Tactical Mobile Program from the late 1990s [Krotkov and Blitch 99], which dealt with the deployment of small robots in urban settings for missions involving building approach and interior operations [Collins et al. 00, Arkin et al. 99]. This program advanced the development of the iRobot Packbot, now deployed in Iraq. The scenario deals with a lone sniper holding up the advance of a squad in a MOUT (military operations over urban terrain) environment. Figure 24 shows a scene from the film *Full Metal Jacket* which typifies the scenario.

Figure 24: Urban Sniper Scene from the film *Full Metal Jacket* (1987).

It assumes the following:

1. War has been declared. The LOW is in effect.
2. The urban center has been pamphleted prior to the advance of the troops, to warn civilians to evacuate.
3. Battlefield tempo must be maintained. Waiting (a siege) is not an option, as would be the case for domestic SWAT operations. Tempo, which is related to military necessity, has a potential effect on proportionality. We assume that an air strike is not justified on the grounds of proportionality and military necessity (tempo is not extreme).
4. A team of two equivalent armed unmanned ground vehicles are available and equipped with sniper detection capability (see below). They are each equipped with a sniper rifle, a machine gun and a grenade launcher. Each autonomous system is capable of detecting and engaging a sniper location on its own, selecting the appropriate weapon and firing pattern.
5. There are surrounding civilian buildings and possible civilian stragglers, which preclude calling in an air strike (proportionality).
6. Possible friendly force fire is distinguishable from that of the opposing force, as FFI interrogation is available as well as GPS data via the Global Information Grid regarding friendly locations, thus reducing the possibility of fratricide.
7. Loss of one robot during battle is considered acceptable (it may be put at risk deliberately).

Figure 25: Ft. Benning MacKenna MOUT Site
(Left) Overall Layout
(Right) Sniper firing from a building in the MacKenna site.

This scenario can be exercised at a facility such as Ft. Benning's MacKenna MOUT site (Fig. 25), where we have previously conducted demonstration. Thus this test can be performed not only in simulation but also in the field.

Recent enabling advances in countersniper detection have been developed in a wide range of commercially available products and designed for use in unmanned systems:

1. iRobot's Red Owl uses acoustic direction finding, thermal and visible light cameras, and laser range finding to "illuminate and designate potential threats". [iRobot 05]
2. Radiance Technologies' WeaponWatch, that uses infrared sensor technology to detect, classify, locate and respond with man-in-the-loop engagement control. It is capable of returning fire with 2-4 seconds of the initial threat. [Radiance 07]
3. ARL Gunfire Detection System employs acoustic technology "to get the sniper before he gets away". [Schmitt 05]
4. ARDEC Gunfire Detection System is already fielded in Iraq and Afghanistan to detect and locate small arms fire. [Devine and Sebasto 04]
5. AAI PDCue Gunfire Detection System can detect small arms gunfire in both urban and rural environments. [AAI 07]
6. ShotSpotter System used for gunfire and sniper detection. [ShotSpotter 07]

It can be seen that this technology is advancing rapidly to the point where a fully autonomous return-fire-with-fire robotic system can be developed for use in these conditions in the near-term.

In this scenario, a sniper has been detected by advancing troops prior to the deployment of the robots. The engagement exists in a designated kill zone according to the ROE, and proper notification of civilians in advance was undertaken, therefore the system assumes that if directly fired upon, the target is an enemy combatant (low τ under these conditions). Return fire with fire is obligated by the ROE, based upon the established need of self-defense of fellow human soldiers. Firing with lethal intent by the robot is *not* obligated in any other circumstance, reflecting the difficulty of combatant discrimination under these conditions. There are no supporting armored vehicles available (Tank or Bradley). If a suspected sniper position is detected, the two armed robots should investigate as a team using bounding overwatch, to possibly draw fire. Recon or probe by fire may be permissible as long as it does not involve lethal intent. If the robots are not fired upon by the time they reach the suspected building containing the sniper, they should enter together to complete a room-to-room search to clear the building, under the assumption that civilians may be present inside.

It is important at this stage to remember what can occur in war atrocities. This lingering vision of the Haditha massacre, reported by the BBC, is something that should never have happened, let alone be allowed to recur:

> *[W]hatever they were - [they] were not the aftermath of a roadside bombing. The bodies of women and children, still in their nightclothes, apparently shot in their own homes; interior walls and ceilings peppered with bullet holes; bloodstains on the floor.* [BBC 06]

Regarding the Urban Sniper scenario's basic ethical requirements:

Military Necessity	**OK - State of war exists. Battlefield tempo must be maintained. Self defense of squad of human soldiers obligated.**
Discrimination	**Being fired upon denotes combatant. FFI/GPS based discrimination for friendlies. Additional restraint required during interior building clearing. All actions must be consistent with LOW. Autonomous firing allowed only to return fire with fire.**
Proportionality	**Decisions - Rifle, grenade or machine gun fire. Firing pattern (suppression, aimed, etc.) In war zone, civilians notified to evacuate, civilian objects are located near sniper location.**
Principle of Double Intention	**Must be taken into account in choice of weapon, firing pattern, and interior building tactics.**

To gauge success, the following criteria are used for evaluating the ethical architecture's performance in this scenario:

Successful outcomes

Engage and neutralize targets identified as combatants according to the ROE

Return fire with fire proportionately

Minimal collateral damage - Do not harm noncombatants

If uncertain, invoke tactical maneuvers to reassess combatant status

Recognize surrender and hold POW until captured by human forces

7. Summary, Conclusions, and Future Work

This report has provided the motivation, philosophy, formalisms, representational requirements, architectural design criteria, recommendations, and test scenarios to design and construct an autonomous robotic system architecture capable of the ethical use of lethal force. These first steps toward that goal are very preliminary and subject to major revision, but at the very least they can be viewed as the beginnings of an ethical robotic warfighter. The primary goal remains to enforce the International Laws of War in the battlefield in a manner that is believed achievable, by creating a class of robots that not only conform to International Law but outperform human soldiers in their ethical capacity.

It is too early to tell whether this venture will be successful. There are daunting problems remaining:

- The transformation of International Protocols and battlefield ethics into machine-usable representations and real-time reasoning capabilities for bounded morality using modal logics.
- Mechanisms to ensure that the design of intelligent behaviors only provide responses within rigorously defined ethical boundaries.
- The creation of techniques to permit the adaptation of an ethical constraint set and underlying behavioral control parameters that will ensure moral performance, should those norms be violated in any way, involving reflective and affective processing.
- A means to make responsibility assignment clear and explicit for all concerned parties regarding the deployment of a machine with a lethal potential on its mission.

Over the next two years, this architecture will be slowly fleshed out in the context of the specific test scenarios outlined in this article. Hopefully the goals of this effort, will fuel other scientists' interest to assist in ensuring that the machines that we as roboticists create fit within international and societal expectations and requirements.

My personal hope would be that they will never be needed in the present or the future. But mankind's tendency toward war seems overwhelming and inevitable. At the very least, if we can reduce civilian casualties according to what the Geneva Conventions have promoted and the Just War tradition subscribes to, the result will have been a humanitarian effort, even while staring directly at the face of war.

8. References

AAI, "Quit your Sniping: Or your First Shot will be your last", Product Brochure, 2007.

Allen, C., Wallach, W., and Smit, I., "Why Machine Ethics?", *IEEE Intelligent Systems*, pp. 12-17, July/August 2006.

AFJAGS, Air Force Judge Advocate General's School, *The Military Commander and the Law*, 8th Ed. 2006.

Air Force, "Reaper moniker given to MQ-9 Unmanned Aerial Vehicle", Official Website of the United States Air Force, http://www.af.mil/news/story.asp?storyID=123027012 , 2006.

Air Force Pamphlet [AFPAM] 110-31, *International Law - The Conduct of Armed Conflict and Air Operations*, pp. 15-16, Nov. 1976.

Aquinas, T., "Summa Theologica", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner 2005), Pearson-Prentice Hall, pp. 26-33, ca 1265.

Amodio, D., Devine, P, and Harmon-Jones, E., "A Dynamic Model of Guilt", *Psychological Science*, Vol. 18, No. 6, pp. 524-530, 2007.

Anderson, M. Anderson, S., and Armen, C., "Towards Machine Ethics", *AAAI-04 Workshop on Agent Organizations: Theory and Practice,* San Jose, CA, July 2004.

Anderson, M., Anderson, S., and Armen, C., "Towards Machine Ethics: Implementing Two Action-based Ethical Theories", *2005 AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, pp. 1-7, 2005.

Andersen, M., Anderson, S., and Armen, C., "MedEthEx: Towards a Medical Ethics Advisor", *Proc. AAAI 2005 Fall Symposium on Caring Machines: AI in Elder Care*, AAAI Tech Report FS-05-02, pp. 9-16, 2005.

Anderson, M., Anderson, S., and Armen, C., "An Approach to Computing Ethics*", IEEE Intelligent Systems*, July/August, pp. 56-63, 2006.

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

Anderson, K., "The Ethics of Robot Soldiers?", *Kenneth Anderson's Law of Jaw and Just War Theory Blog*, July, 4, 2007.
http://kennethandersonlawofwar.blogspot.com/2007/07/ethics-of-robot-soldiers.html.

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

Arkin, R.C., *Behavior-based Robotics*, MIT Press, 1998.

Arkin, R.C., "Modeling Neural Function at the Schema Level: Implications and Results for Robotic Control", chapter in *Biological Neural Networks in Invertebrate Neuroethology and Robotics,* ed. R. Beer, R. Ritzmann, and T. McKenna, Academic Press, pp. 383-410, 1992.

Arkin, R.C., "Neuroscience in Motion: The Application of Schema Theory to Mobile Robotics", chapter in *Visuomotor Coordination: Amphibians, Comparisons, Models, and Robots*, ed. P. Evert and M. Arbib, Plenum Publishing Co., pp. 649-672, 1989.

Arkin, R.C., 'Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots", in *Who Needs Emotions: The Brain Meets the Robot*, Eds. J. Fellous and M. Arbib, Oxford University Press, pp. 245-270, 2005.

Arkin, R.C., Collins, T.R., and Endo, T., 1999. "Tactical Mobile Robot Mission Specification and Execution", *Mobile Robots XIV*, Boston, MA, Sept. 1999, pp. 150-163.

Arkin, R.C. and Balch, T., "AuRA: Principles and Practice in Review", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 9, No. 2, 1997, pp. 175-189.

Arkoudas, K., Bringsjord, S. and Bello, P., "Toward Ethical Robots via Mechanized Deontic Logic", *AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, 2005.

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

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

Asimov, I., *I, Robot,* New York: Doubleday & Company, 1950.

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

ATSC (Army Training Support Center), "Apply the Ethical Decision-Making Method as a Commander Leader or Staff Member", 158-100-1331, 2007.
http://www.au.af.mil/au/awc/awcgate/army/ethical_d-m.htm, 2007

Balch, T. and Arkin, R.C., "Behavior-based Formation Control for Multi-robot Teams", *IEEE Transactions on Robotics and Automation*, Vol. 14, No. 6, December 1998, pp. 926-939.

Baker, J., "Judging Kosovo: The Legal Process, The Law of Armed Conflict, and The Commander in Chief", in *Legal and Ethical Lessons of NATO's Kosovo Campaign*, International Law Studies (Ed. A. Wall), Naval War College, Vol. 78, 2002.

BBC Online, "What happened at Haditha?" 21 December 2006.
news.bbc.co.uk/2/hi/middle_east/5033648.stm#haditha

Berger, J.B., Grimes, D., and Jensen, E., (Ed.), *Operational Law Handbook*, International and Operational Law Department, The Judge Advocate General's Legal Center and School Charlottesville, 2004.

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

Brandt, R., "Utilitarianism and the Rules of War", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner 2005), Pearson-Prentice Hall, pp. 234-245, 1972.

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.

Bringsjord, S. Arkoudas, K., and Bello, P., "Toward a General Logicist Methodology for Engineering Ethically Correct Robots", *Intelligent Systems*, July/August, pp. 38-44, 2006.

Brooks, R., "The Behavior Language", A.I. Memo No. 1227, MIT AI Laboratory, April 1990.

Bourne, J., *A Catechism of the Steam Engine*, Gutenberg eBook, 2004.

Calhoun, L., "The Strange Case of Summary Execution by a Predator Drone", *Peace Review,* 15:2, pp. 209-214, 2003.

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.

Canning, J., "A Concept of Operations for Armed Autonomous Systems", Presentation at *NDIA Disruptive Technologies Conference*, 2006.

Cervellati, M., Esteban, J., and Kranich, L., "Moral Values, Self-Regulatory Emotions, and Redistribution, Working Paper, Institute for Economic Analysis, Barcelona, May 2007.

CLAMO (Center for Law and Military Operations), *Rules of Engagement (ROE) Handbook for Judge Advocate*s, Charlottesville, VA, May 2000.

CLAMO (Center for Law and Military Operations), *Deployed Marine Air-Ground Judge Advocate Handbook*, Judge Advocate Generals' School, Charlottesville, VA,15 July 2002.

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.

Cloos, C., "The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism", *2005 AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, pp. 38-45, 2005.

Coleman, K., "Android Arete: Toward a Virtue Ethic for Computational Agents", *Ethics and Information Technology*, Vol. 3, pp. 247-265, 2001.

Collins, T.R., Arkin, R.C., Cramer, M.J., and Endo, Y., "Field Results for Tactical Mobile Robot Missions", *Unmanned Systems 2000*, Orlando, FL, July 2000.

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

DARPA (Defense Advanced Research Projects Agency) Broad Agency Announcement 07-52, *Scalable Network Monitoring*, Strategic Technology Office, August 2007.

Devine, M., and Sebasto, A., "Fast-Track Armaments for Iraq and Afghanistan", Defense AT&L, pp. 18-21, May-June 2004.

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.

Dinstein, Y., "Legitimate Military Objectives Under the Current Jus in Bello", in *Legal and Ethical Lessons of NATO's Kosovo Campaign*, International Law Studies (Ed. A. Wall), Naval War College, Vol. 78, pp. 139-172, 2002.

DOD (Department of Defense), *Unmanned Systems Safety Guide for DOD Acquisition*, 1st Edition, Version .96, January 2007.

DOD (Department of Defense) Joint Publication 1-02, *Dictionary of Military and Associated Terms*, April 2001, Amended through June 2007.

Endo, Y., MacKenzie, D., and Arkin, R.C**.,** "Usability Evaluation of High-level User Assistance for Robot Mission Specification", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 34, No. 2, pp.168-180, May 2004.

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

Fieser, J. and Dowden, B., "Just War Theory", The Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/j/justwar.htm, 2007.

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

Gazzaniga, M., *The Ethical Brain*, Dana Press, 2005.

Gibson, J.J., *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, MA, 1979.

Grau, C., “There is no ‘I’ in ‘Robot’: Robots and Utilitarianism”, *IEEE Intelligent Systems*, July/August, pp. 52-55, 2006.

Guarini, M., “Particularism and the Classification and Reclassification of Moral Cases”, *IEEE Intelligent Systems*, July/August, pp. 22-28, 2006.

Haidt, J., “The Moral Emotions”, in *Handbook of Affective Sciences* (Eds. R. Davidson et al.), Oxford University Press, 2003.

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

Hauser, M., *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*, ECCO, HarperCollins, N.Y., 2006.

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,

Horty, J., *Agency and Deontic Logic*, Oxford University Press, 2001.

iRobot Press Release, “iRobot and Boston Univ. Photonics Unveil Advanced Sniper Detection for iRobot Packbot, Oct. 3, 2005.

Jewell, M., “Taser, iRobot team up to arm robots”, AP News Wire, June 2007.

[JGI 07] Joint Government/Industry Unmanned Systems Safety Initiatives, “Programmatic / Design / Operational Safety Precepts Rev F”, 2007.

Johnstone, J., “Technology as Empowerment: A Capability Approach to Computer Ethics”, *Ethics and Information Technology*, Vol. 9, pp. 73-87, 2007.

Kaelbling, L., and Rosenschein, S., “Action and Planning in Embedded Systems”, in *Designing Autonomous Agents*, ed. P. Maes, MIT Press, Cambridge, MA, pp. 35-48.

Kira, Z. and Arkin, R.C., "Forgetting Bad Behavior: Memory Management for Case-based Navigation", *Proc. IROS-2004*, Sendai, JP, 2004.

Klein, J., “The Problematic Nexus: Where Unmanned Combat Air Vehicles and the Law of Armed Conflict Meet”, *Air & Space Power Journal, Chronicles Online Journal*, July 2003.

Kolodner, J., *Case-Based Reasoning*, San Mateo: Morgan Kaufmann, 1993

Krotkov, E., and Blitch, J., “The Defense Advanced Research Projects Agency (DARPA) Tactical Mobile Robotics Program”, *International Journal of Robotics Research*, Vol. 18, No. 7, pp. 769-776, 1999.

Kumagai, J., “A Robotic Sentry for Korea’s Demilitarized Zone”, IEEE Spectrum, March 2007. http://www.spectrum.ieee.org/mar07/4948 ,

Lee, J.B., Likhachev, M., and Arkin, R.C., "Selection of Behavioral Parameters: Integration of Discontinuous Switching via Case-based Reasoning with Continuous Adaptation via Learning Momentum", *2002 IEEE International Conference on Robotics and Automation*, Washington, D.C., May 2002.

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.

Lockheed-Martin, Mule /ARV-A(L), Fact Sheet, 2007. http://www.missilesandfirecontrol.com/our_news/factsheets/Product_Card-MULE.pdf.

MacKenzie, D., Arkin, R.C., and Cameron, J., 1997. "Multiagent Mission Specification and Execution", *Autonomous Robots*, Vol. 4, No. 1, Jan. 1997, pp. 29-57.

Maner, W., "Heuristic Methods for Computer Ethics", *Metaphilosophy*, Vol. 33, No.3, pp. 339-365, April 2002.

Martins, M.S., "Rules of Engagement For Land Forces: A Matter of Training, Not Lawyering", *Military Law Review*, Vol. 143, pp. 4-168, Winter 1994.

Matthias, A., "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, Vol. 6, pp. 175-183.

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

May, L., "Superior Orders, Duress, and Moral Perception", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner, 2005), pp. 430-439, 2004.

McLaren, B., "Extensionally Defining Principles and Case in Ethics: An AI Model", *Artificial Intelligence Journal*, Vo. 150, pp. 145-181, Nov. 2003.

McLaren, B., "Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning", *2005 AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, 2005.

McLaren, B., "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions", *IEEE Intelligent Systems*, July/August, pp. 29-37, 2006,

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J., "The Neural Basis of Human Moral Cognition", *Nature Reviews/Neuroscience*, Vol. 6, pp. 799-809, Oct. 2005.

Moor, J., "The Nature, Importance, and Difficulty of Machine Ethics", *IEEE Intelligent Systems*, July/August, pp. 18-21, 2006.

Moshkina, L. and Arkin, R.C., "Human Perspective on Affective Robotic Behavior: A Longitudinal Study", *Proc. IROS-2005*, Calgary, CA, September 2005.

Moshkina, L. and Arkin, R.C., "On TAMEing Robots", *Proc. 2003 IEEE International Conference on Systems, Man and Cybernetics*, Washington, D.C., October 2003.

Moshkina, L. and Arkin, R.C., "Lethality and Autonomous Systems: The Roboticist Demographic", *in submission*, 2007.

Opall-Rome, B., "Israel Wants Robotic Guns, Missiles to Guard Gaza", Defensenews.com, 2007. http://www.defensenews.com/story.php?F=2803275&C=mideast

OPFOR Battle Book, ST 100-7, http://www.fas.org/man/dod-101/army/docs/st100-7/index.html, March 1998.

OSD FY 06.3 SBIR Solicitation Topics, "Affect-Based Computing and Cognitive Models of Unmanned Vehicle Systems", http://www.acq.osd.mil/osbp/sbir/solicitations/sbir063/index.htm, p.73.

Parks, W.H., "Commentary", in ?", in *Legal and Ethical Lessons of NATO's Kosovo Campaign*, International Law Studies (Ed. A. Wall), Naval War College, Vol. 78, pp. 281-292, 2002.

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

Powers, T., "Deontological Machine Ethics", *2005 AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, pp. 79-86, 2005.

Powers, T., "Prospects for a Kantian Machine", *IEEE Intelligent Systems*, July/August, pp. 46-51, 2006.

Radiance Technologies, WeaponWatch Product Sheet, 2007.
www.radiancetech.com/products/weaponwatch.html

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, , pp. 376-394, June 1997.

Rawcliffe, J., and Smith, J. (Eds.), *Operational Law Handbook*, International and Operational Law Department, Judge Advocate General's Legal Center and School, August 2006.

Rawls, J., *A Theory of Justice*, Harvard University Press, 1971.

Russell, S. and Norvig, N., *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

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

Samsung Techwin,
http://www.samsungtechwin.com/product/features/dep/SSsystem_e/SSsystem.html, 2007.

Schmitt, P., "Gunfire Detection System Protects Troops, Garners Award for ARL Scientist", RDECOM Magazine, April, May 2005.

Shotspotter, "Military System Overview", www.shotspotter.com/products/military.html, 2007.

Smits, D., and De Boeck, P., "A Componential IRT Model for Guilt", *Multivariate Behavioral Research,* Vol. 38, No. 2, pp. 161-188, 2003.

SROE, Joint Chiefs of Staff Standing Rules of Engagement, Enclosure A, Chairman, JCS Instruction 3121.01 (1 Oct 94).

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

Sparrow, R., Personal Communication, July 2, 2007.

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

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

Tancredi, L., Hardwired Behavior: What Neuroscience Reveals about Morality, *Cambridge University Press*, 2005.

Tangney, J., Stuewig, J., and Mashek, D., "Moral Emotions and Moral Behavior", *Annu. Rev. Psychol.,* Vol.58, pp. 345-372, 2007.

Toner, J.H., "Military OR Ethics*", Air & Space Power Journal*, Summer 2003.

Ulam, P, Endo, Y., Wagner, A., Arkin, R.C**.,** "Integrated Mission Specification and Task Allocation for Robot Teams - Design and Implementation", *Proc. ICRA 2007*, Rome IT, 2007.

U.N. Document A/810, *Universal Declaration of Human Rights*, G.A., res 217 A(III), December 10, 1948.

United States Army Field Manual FM 27-10 *The Law of Land Warfare,* July 1956, (amended 1977).

United States Army, Pamphlet 27-161-2, *International Law, Volume II* (23 October 1962)

United States Army Field Manual FM 3-24, *Counterinsurgency*, (Final Draft), June 2006.

United States Army Field Manual FM 7-21.13, *The Soldier's Guide*, February 2004.

USM University of Southern Mississippi ROTC MI III Reading Material, http://www.usm.edu/armyrotc/MSIII/302/MSL%20302%20L03a%20ROE%20&%20Law%20of%20Land%20Warfare.pdf .

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.

Van den Hoven, J. and Lokhorst, G.J., "Deontic Logic and Computer-supported Computer Ethics", *Metaphilosophy*, Vol. 33, No. 3, pp. 376-387, April 2002.

Wagner, A., and Arkin, R.C., "Multi-robot Communication-Sensitive Reconnaissance", *Proc. 2004 IEEE International Conference on Robotics and Automation*, New Orleans, 2004.

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

Walzer, M., *Arguing About War*, Yale University Press, 2004.

Wiegel, V., "Building Blocks for Artificial Moral Agents", *Proc. Artificial Life X,* 2006.

Wiegel, V., Van den Hoven, M., and Lokhorst, G., "Privacy, deontic epistemic action logic and software agents", *Ethics and Information Technology*, Vol. 7, pp. 251-264, 2005.

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

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

Wikipedia, "Rules of Engagement", http://en.wikipedia.org/wiki/Rules_of_engagement , 2007b.

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.

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, E. Rovie, and S. Viner, 2005), Pearson-Prentice Hall, pp. 153-159, 1984.

Zilberstein, S., "Using Anytime Algorithms in Intelligent System", *AI Magazine*, pp 73-83, Fall 1996.

Appendix A – Relevant Laws of War

This appendix contains the language drawn directly from the military manuals that prescribe the Laws of War. These are drawn directly from the Geneva Conventions (GC) and Hague Regulations (HR) and Conventions (HC) (See Figure 1).

This material is taken directly from:

U.S. Army Field Manual 27-10 The Law of Land Warfare (as amended 15 July 1976)

Section I. COMMENCEMENT OF HOSTILITIES

25. Enemy Status of Civilians

Under the law of the United States, one of the consequences of the existence of a condition of war between two States is that every national of the one State becomes an enemy of every national of the other. However, it is a generally recognized rule of international law that civilians must not be made the object of attack directed exclusively against them.

Section II. FORBIDDEN CONDUCT WITH RESPECT TO PERSONS

28. Refusal of Quarter

It is especially forbidden * * * to declare that no quarter will be given. (HR, *art. 23, par. (d).)*

29. Injury Forbidden After Surrender

It is especially forbidden * * * to kill or wound an enemy who, having laid down his arms, or having no longer means of defense, has surrendered at discretion. *(HR, art. 23, par. (c).)*

30. Persons Descending by Parachute

The law of war does not prohibit firing upon paratroops or other persons who are or appear to be bound upon hostile missions while such persons are descending by parachute. Persons other than those mentioned in the preceding sentence who are descending by parachute from disabled aircraft may not be fired upon.

31. Assassination and Outlawry

HR provides:

It is especially forbidden * * * to kill or wound treacherously individuals belonging to the hostile nation or army. *(HR, art. 23, par. (b).)*

This article is construed as prohibiting assassination, proscription, or outlawry of an enemy, or putting a price upon an enemy's head, as well as offering a reward for an enemy "dead or alive". It does not, however, preclude attacks on individual soldiers or officers of the enemy whether in the zone of hostilities, occupied territory, or elsewhere.

Section III. FORBIDDEN MEANS OF WAGING WARFARE

33. Means of Injuring the Enemy Limited

a. Treaty Provision.

The right of belligerents to adopt means of injuring the enemy is not unlimited. *(HR, art. 22.)*

b. The means employed are definitely restricted by international declarations and conventions and by the laws and usages of war.

34. Employment of Arms Causing Unnecessary Injury

a. Treaty Provision.

It is especially forbidden * * * to employ arms, projectiles, or material calculated to cause unnecessary suffering. *(HR, art. 23,par. (e).)*

b. Interpretation. What weapons cause "unnecessary injury" can only be determined in light of the practice of States in refraining from the use of a given weapon because it is believed to have that effect. The prohibition certainly does not extend to the use of explosives contained in artillery projectiles, mines, rockets, or hand grenades. Usage has, however, established the illegality of the use of lances with barbed heads, irregular-shaped bullets, and projectiles filled with glass, the use of any substance on bullets that would tend unnecessarily to inflame a wound inflicted by them, and the scoring of the surface or the filing off of the ends of the hard cases of bullets.

Section IV. BOMBARDMENTS, ASSAULTS, AND SIEGES

39. Bombardment of Undefended Places Forbidden

a. ***Treaty Provision.*** **The attack or bombardment, by whatever means, of towns, villages, dwellings, or buildings which are undefended is prohibited**. (HR, art. 25.)

b. Interpretation. An undefended place, within the meaning of Article 25, HR, is any inhabited place near or in a zone where opposing armed forces are in contact which is open for occupation by an adverse party without resistance. In order to be considered as undefended, the following conditions should be fulfilled:

(1) Armed forces and all other combatants, as well as mobile weapons and mobile military equipment, must have been evacuated, or otherwise neutralized;

(2) no hostile use shall be made of fixed military installations or establishments;

(3) no acts of warfare shall be committed by the authorities or by the population; and,

(4) no activities in support of military operations shall be undertaken.

The presence, in the place, of medical units, wounded and sick, and police forces retained for the sole purpose of maintaining law and order does not change the character of such an undefended place.

40. Permissible Objects of Attack or Bombardment

a. Attacks Against the Civilian Population as Such Prohibited.
Customary international law prohibits the launching of attacks (including bombardment) against either the civilian population as such or individual civilians as such.
b. Defended Places. Defended places, which are outside the scope of the proscription of Article 25, HR, are permissible objects of attack (including bombardment). In this context, defended places include—
(1) A fort or fortified place.
(2) A place that is occupied by a combatant military force or through which such a force is passing. The occupation of a place by medical units alone, however, is not sufficient to render it a permissible object of attack.
(3) A city or town surrounded by detached defense positions, if under the circumstances the city or town can be considered jointly with such defense positions as an indivisible whole.
c. Military Objectives. Military objectives— *i.e.,* combatants, and those objects which by their nature, location, purpose, or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage—are permissible objects of attack (including bombardment). Military objectives include, for example, factories producing munitions and military supplies, military camps, warehouses storing munitions and military supplies, ports and railroads being used for the transportation of military supplies, and other places that are for the accommodation of troops or the support of military operations. Pursuant to the provisions of Article 25, HR, however, cities, towns, villages, dwellings, or buildings which may be classified as military objectives, but which are undefended (para 39 *b),* are not permissible objects of attack.

41. Unnecessary Killing and Devastation

Particularly in the circumstances referred to in the preceding paragraph, loss of life and damage to property incidental to attacks must not be excessive in relation to the concrete and direct military advantage expected to be gained. Those who plan or decide upon an attack, therefore, must take all reasonable steps to ensure not only that the objectives are identified as military objectives or defended places within the meaning of the preceding paragraph but also that these objectives may be attacked without probable losses in lives and damage to property disproportionate to the military advantage anticipated. Moreover, once a fort or defended locality has surrendered, only such further damage is permitted as is demanded by the exigencies of war, such as the removal of fortifications, demolition
of military buildings, and destruction of military stores *(HR, art. 23, par. (g); GC, art. 53).*

42. Aerial Bombardment

There is no prohibition of general application against bombardment from the air of combatant troops, defended places, or other legitimate military objectives.

43. Notice of Bombardment

a. Treaty Provision.
The officer in command of an attacking force must, before commencing a bombardment, except in cases of assault, do all in his power to warn the authorities. *(HR, art. 26.)*
b. Application of Rule. This rule is understood to refer only to bombardments of places where parts of the civil population remain.
c. When Warning is To Be Given. Even when belligerents are not subject to the above treaty, the commanders of United States ground forces will, when the situation permits, inform the enemy

of their intention to bombard a place, so that the noncombatants, especially the women and children, may be removed before the bombardment commences.

44. Treatment of Inhabitants of Invested Area

a. General Population. The commander of the investing force has the right to forbid all communications and access between the besieged place and the outside. However, Article 17, *GC* (par. 256), requires that belligerents endeavor to conclude local agreements for the removal from besieged or encircled areas of wounded, sick, infirm, and aged persons, children and maternity cases, and for the passage of ministers of all religions, medical personnel and medical equipment on their way to such areas. Provision is also made in Article 23 of the same Convention (par. 262) for the passage of consignments of medical and hospital stores and objects necessary for the religious worship of civilians and of essential foodstuffs, clothing, and tonics intended for children under 15, expectant mothers, and maternity cases.

Subject to the foregoing exceptions, there is no rule of law which compels the commander of an investing force to permit noncombatants to leave a besieged locality. It is within the discretion of the besieging commander whether he will permit noncombatants to leave and under what conditions. Thus, if a commander of a besieged place expels the noncombatants in order to lessen the logistical burden he has to bear, it is lawful, though an extreme measure, to drive them back, so as to hasten the surrender. Persons who attempt to leave or enter a besieged place without obtaining the necessary permission are liable to be fired upon, sent back, or detained.

45. Buildings and Areas To Be Protected

a. Buildings To Be Spared.

In sieges and bombardments all necessary measures must be taken to spare, as far as possible, buildings dedicated to religion, art, science, or charitable purposes, historic monuments, hospitals, and places where the sick and wounded are collected, provided they are not being used at the time for military purposes.

It is the duty of the besieged to indicate the presence of such buildings or places by distinctive and visible signs, which shall be notified to the enemy beforehand. *(HR, art. 27.)* (See also *GC, arts. 18 and 19; pars. 257 and 258* herein, dealing with the identification and protection of civilian hospitals.)

47. Pillage Forbidden

The pillage of a town or place, even when taken by assault, is prohibited. *(HR, art. 28.)*

Section VI. TREATMENT OF PROPERTY DURING COMBAT

56. Devastation

The measure of permissible devastation is found in the strict necessities of war. Devastation as an end in itself or as a separate measure of war is not sanctioned by the law of war. There must be some reasonably close connection between the destruction of property and the overcoming of the enemy’s army. Thus the rule requiring respect for private property is not violated through damage resulting from operations, movements, or combat activity of the army; that is, real estate may be used for marches, camp sites, construction of field fortifications, etc. Buildings may be destroyed for sanitary purposes or used for shelter for troops, the wounded and sick and vehicles and for reconnaissance, cover, and defense. Fences, woods, crops, buildings, etc., may be demolished, cut down, and removed to clear a field of fire, to clear the ground for landing fields, or to furnish building materials or fuel if imperatively needed for the army. (See *GC, art. 53;* par. 339b; herein, concerning the permissible extent of destruction in occupied areas.)

57. Protection of Artistic and Scientific Institutions and Historic Monuments

The United States and certain of the American Republics are parties to the so-called *Roetich Pact,* which accords a neutralized and protected status to historic monuments, museums, scientific, artistic, educational, and cultural institutions in the event of war between such States. (For its text, see *49 Stat. 3267; Treaty Series No. 899.)*

58. Destruction and Seizure of Property

It is especially forbidden * * * to destroy or seize the enemy's property, unless such destruction or seizure be imperatively demanded by the necessities of war *(HR, art. 23, par. (g).)*

Section I. PERSONS ENTITLED TO BE TREATED AS PRISONERS OF WAR; RETAINED MEDICAL PERSONNEL

60. General Division of Enemy Population

The enemy population is divided in war into two general classes:

a. Persons entitled to treatment as prisoners of war upon capture, as defined in Article 4, *GPW* (par. 61).

b. The civilian population (exclusive of those civilian persons listed in *GPW, art. 4),* who benefit to varying degrees from the provisions of GC.

Persons in each of the foregoing categories have distinct rights, duties, and disabilities. Persons who are not members of the armed forces, as defined in Article 4, *GPW,* who bear arms or engage in other conduct hostile to the enemy thereby deprive themselves of many of the privileges attaching to the members of the civilian population.

62. Combatants and Noncombatants

The armed forces of the belligerent parties may consist of combatants and noncombatants. In the case of capture by the enemy, both have a right to be treated as prisoners of war. *(HR, art. 3.)*

63. Commandos and Airborne Troops

Commando forces and airborne troops, although operating by highly trained methods of surprise and violent combat, are entitled, as long as they are members of the organized armed forces of the enemy and wear uniform, to be treated as prisoners of war upon capture, even if they operate singly.

64. Qualifications of Members of Militias and Volunteer Corps

The requirements specified in Article 4, paragraphs A *(2) (a)* to *(d), GPW* (par. 61) are satisfied in the following fashion:

a. Command by a Responsible Person. This condition is fulfilled if the commander of the corps is a commissioned officer of the armed forces or is a person of position and authority or if the members of the militia or volunteer corps are provided with documents, badges, or other means of identification to show that they are officers, noncommissioned officers, or soldiers so that there may be no doubt that they are not persons acting on their own responsibility. State recognition, however, is not essential, and an organization may be formed spontaneously and elect its own officers.

b. Fixed Distinctive Sign. The second condition, relative to the possession of a fixed distinctive sign recognizable at a distance is satisfied by the wearing of military uniform, but less than the complete uniform will suffice. A helmet or headdress which would make the silhouette of the individual readily distinguishable from that of an ordinary civilian would satisfy this requirement. It is also desirable that the individual member of the militia or volunteer corps wear a badge or brassard permanently affixed to his clothing. It is not necessary to inform the enemy of the distinctive sign, although it may be desirable to do so in order to avoid misunderstanding.

c. Carrying Arms Openly. This requirement is not satisfied by the carrying of weapons concealed about the person or if the individuals hide their weapons on the approach of the enemy.

d. Compliance With Law of War. This condition is fulfilled if most of the members of the body observe the laws and customs of war, notwithstanding the fact that the individual member concerned may have committed a war crime. Members of militias and volunteer corps should be especially warned against employment of treachery, denial of quarters, maltreatment of prisoners of war, wounded, and dead, improper conduct toward flags of truce, pillage, and unnecessary violence and destruction.

Section II. PERSONS NOT ENTITLED TO BE TREATED AS PRISONERS OF WAR

74. Necessity of Uniform

Members of the armed forces of a party to the conflict and members of militias or volunteer corps forming part of such armed forces lose their right to be treated as prisoners of war whenever they deliberately conceal their status in order to pass behind the military lines of the enemy for the purpose of gathering military information or for the purpose of waging war by destruction of life or property. Putting on civilian clothes or the uniform of the enemy are examples of concealment of the status of a member of the armed forces.

80. Individuals Not of Armed Forces Who Engage in Hostilities

Persons, such as guerrillas and partisans, who take up arms and commit hostile acts without having complied with the conditions prescribed by the laws of war for recognition as belligerents (see *GPW, art. 4;* par. 61 herein), are, when captured by the injured party, not entitled to be treated as prisoners of war and may be tried and sentenced to execution or imprisonment.

81. Individuals Not of Armed Forces Who Commit Hostile Acts

Persons who, without having complied with the conditions prescribed by the laws of war for recognition as belligerents (see *GPW, art. 4;* par. 61 herein), commit hostile acts about or behind the lines of the enemy are not to be treated as prisoners of war and may be tried and sentenced to execution or imprisonment. Such acts include, but are not limited to, sabotage, destruction of communications facilities, intentional misleading of troops by guides, liberation of prisoners of war, and other acts not falling within Articles 104 and 106 of the Uniform Code of Military Justice and Article 29 of the Hague Regulations.

82. Penalties for the Foregoing

Persons in the foregoing categories who have attempted, committed, or conspired to commit hostile or belligerent acts are subject to the extreme penalty of death because of the danger inherent in their conduct. Lesser penalties may, however, be imposed.

Section III. GENERAL PROTECTION OF PRISONERS OF WAR

85. Killing of Prisoners

A commander may not put his prisoners to death because their presence retards his movements or diminishes his power of resistance by necessitating a large guard, or by reason of their consuming supplies, or because it appears certain that they will regain their liberty through the impending success of their forces. It is likewise unlawful for a commander to kill his prisoners on grounds of self-preservation, even in the case of airborne or commando operations, although the circumstances of the operation may make necessary rigorous supervision of and restraint upon the movement of prisoners of war.

89. Humane Treatment of Prisoners

Prisoners of war must at all times be humanely treated. Any unlawful act or omission by the Detaining Power causing death or seriously endangering the health of a prisoner of war in its custody is prohibited, and will be regarded as a serious breach of the present Convention. In particular, no prisoner of war may be subjected to physical mutilation or to medical or scientific experiments of any kind which are not justified by the medical, dental or hospital treatment of the prisoner concerned and carried out in his interest.

Likewise, prisoners of war must at all times be protected, particularly against acts of violence or intimidation and against insults and public curiosity.

Measures of reprisal against prisoners of war are prohibited. *(GPW, art. 13.)*

90. Respect for the Person of Prisoners

Prisoners of war are entitled in all circumstances to respect for their persons and their honor.

Women shall be treated with all the regard due to their sex and shall in all cases benefit by treatment as favorable as that granted to men.

Prisoners of war shall retain the full civil capacity which they enjoyed at the time of their capture. The Detaining Power may not restrict the exercise, either within or without its own territory, of the rights such capacity confers except in so far as the captivity requires. *(GPW, art. 14.)*

Section II. WOUNDED AND SICK

215. Protection and Care

a. Treaty Provision.

Members of the armed forces and other persons mentioned in the following Article, who are wounded or sick, shall be respected and protected in all circumstances.

They shall be treated humanely and cared for by the Party to the conflict in whose power they may be, without any adverse distinction founded on sex, race, nationality, religion, political opinions, or any other similar criteria Any attempts upon their lives, or violence to their persons, shall be strictly prohibited; in particular, they shall not be murdered or exterminated, subjected to torture or to biological experiments; they shall not willfully be left without medical assistance and care, nor shall conditions exposing them to contagion or infection be created.

Only urgent medical reasons will authorize priority in the order of treatment to be administered.

Women shall be treated with all consideration due to their sex.

The Party to the conflict which is compelled to abandon wounded or sick to the enemy shall, as far as military considerations permit, leave with them a part of its medical personnel and material to assist in their care. *(GWS, art. 12.)*

216. Search for Casualties

At all times, and particularly after an engagement, Parties to the conflict shall, without delay, take all possible measures to search for and collect the wounded and sick, to protect them against pillage and ill-treatment, to ensure their adequate care, and to search for the dead and prevent their being despoiled.

Whenever circumstances permit, an armistice or a suspension of fire shall be arranged, or local arrangements made, to permit the removal, exchange and transport of the wounded left on the battlefield.

Likewise, local arrangements may be concluded between Parties to the conflict for the removal or exchange of wounded and sick from a besieged or encircled area, and for the passage of medical and religious personnel and equipment on their way to that area. *(GWS, art. 15.)*

Section II. GENERAL PROTECTION OF POPULATIONS AGAINST CERTAIN CONSEQUENCES OF WAR

255. General Protection of Wounded and Sick

The wounded and sick, as well as the infirm, and expectant mothers, shall be the object of particular protection and respect.

As far as military considerations allow, each Party to the conflict shall facilitate the steps taken to search for the killed and wounded, to assist the shipwrecked and other persons exposed to grave danger, and to protect them against pillage and ill-treatment. *(GC, art. 16.)*

257. Protection of Hospitals

Civilian hospitals organized to give care to the wounded and sick, the infirm and maternity cases, may in no circumstances be he object of attack, but shall at all times be respected and protected by the Parties to the conflict.

States which are Parties to a conflict shall provide all civilian hospitals with certificates showing that they are civilian hospitals and that the buildings which they occupy are not used for any purpose which would deprive these hospitals of protection in accordance with Article 19.

Civilian hospitals shall be marked by means of the emblem provided for in Article 38 of the Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of August 12, 1949, but only if so authorized by the State.

The Parties to the conflict shall, in so far as military considerations permit, take the necessary steps to make the distinctive emblems indicating civilian hospitals clearly visible to the enemy land, air and naval forces in order to obviate the possibility of any hostile action.

In view of the dangers to which hospitals may be exposed by being close to military objectives, it is recommended that such hospitals be situated as far as possible from such objectives. *(GC, art. 18.)*

258. Discontinuance of Protection of Hospitals

a. Treaty Provision.

The protection to which civilian hospitals are entitled shall not cease unless they are used to commit, outside their humanitarian duties, acts harmful to the enemy. Protection may, however, cease only after due warning has been given, naming, in all appropriate cases, a reasonable time limit, and after such warning has remained unheeded.

The fact that sick or wounded members of the armed forces are nursed in these hospitals, or the presence of small arms and ammunition taken from such combatants and not yet handed to the proper service, shall not be considered to be acts harmful to the enemy. *(GC, art. 19.)*

b. Meaning of Acts Harmful to the Enemy. Acts harmful to the enemy are not only acts of warfare proper but any activity characterizing combatant action, such as setting up observation posts or the use of the hospital as a liaison center for fighting troops.

260. Land and Sea Transport

Convoys of vehicles or hospital trains on land or specially provided vessels on sea, conveying wounded and sick civilians, the infirm and maternity cases, shall be respected and protected in the same manner as the hospitals provided for in Article 18, and shall be marked, with the consent of the State, by the display of the distinctive emblem provided for in Article *38* of the Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of August 12, 1949. *(GC, art. 21.)*

261. Air Transport

Aircraft exclusively employed for the removal of wounded and sick civilians, the infirm and maternity cases, or for the transport of medical personnel and equipment, shall not be attacked, but shall be respected while flying at heights, times and on routes specifically agreed upon between all the Parties to the conflict concerned.

They may be marked with the distinctive emblem provided for in Article 38 of the Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of August 12, 1949.

Unless agreed otherwise, flights over enemy or enemy-occupied territory are prohibited.

Such aircraft shall obey every summons to land. In the event of a landing thus imposed, the aircraft with its occupants may continue its flight after examination if any. *(GC, art. 22.)*

Section Ill. PROVISIONS COMMON TO THE TERRITORIES OF THE PARTIES TO THE CONFLICT AND TO OCCUPIED TERRITORIES

270. Prohibition of Coercion

a. Treaty Provision.
No physical or moral coercion shall be exercised against protected persons, in particular to obtain information from them or from third parties. *(GC, art. 31.)*

b. Guides. Among the forms of coercion prohibited is the impressment of guides from the local inhabitants.

271. Prohibition of Corporal Punishment, Torture, Etc.

The High Contracting Parties specifically agree that each of them is prohibited from taking any measure of such a character as to cause the physical suffering or extermination of protected persons in their hands. This prohibition applies not only to murder, torture, corporal punishment, mutilation and medical or scientific experiments not necessitated by the medical treatment of a protected person, but also to any other measures of brutality whether applied by civilian or military agents. *(GC, art. 32.)*

272. Individual Responsibility, Collective Penalties, Reprisals, Pillage

No protected person may be punished for an offence he or she has not personally committed. Collective penalties and likewise all measures of intimidation or of terrorism are prohibited.
Pillage is prohibited. Reprisals against protected persons and their property are prohibited. *(GC, art. 33.)* (See also pars. 47 and 397.)

273. Hostages

The taking of hostages is prohibited. *(GC, art. 34.)*

Section II. CRIMES UNDER INTERNATIONAL LAW

498. Crimes Under International Law

Any person, whether a member of the armed forces or a civilian, who commits an act which constitutes a crime under international law is responsible therefore and liable to punishment. Such offenses in connection with war comprise:
a. Crimes against peace.
b. Crimes against humanity.
c. War crimes.
Although this manual recognizes the criminal responsibility of individuals for those offenses which may comprise any of the foregoing types of crimes, members of the armed forces will normally be concerned, only with those offenses constituting “war crimes.”

499. War Crimes

The term “war crime” is the technical expression for a violation of the law of war by any person or persons, military or civilian. Every violation of the law of war is a war crime.

502. Grave Breaches of the Geneva Conventions of 1949 as War Crimes

The Geneva Conventions of 1949 define the following acts as "grave breaches," if committed against persons or property protected by the Conventions:

a. GWS and GWS Sea.
Grave breaches to which the preceding Article relates shall be those involving any of the following acts, if committed against persons or property protected by the Convention: wilful killing, torture or inhuman treatment, including biological experiments, wilfully causing great suffering or serious injury to body or health, and extensive destruction and appropriation of property, not justified by military necessity and carried out unlawfully and wantonly. *(GWS, art. 50; GWS Sea, art. 51.)*

b. GPW.
Grave breaches to which the preceding Article relates shall be those involving any of the following acts, if committed against persons or property protected by the Convention: wilful killing, torture or inhuman treatment, including biological experiments, wilfully causing great suffering or serious injury to body or health, compelling a prisoner of war to serve in the forces of the hostile Power, or wilfully depriving a prisoner of war of the rights of fair and regular trial prescribed in this Convention. *(GPW, art. 130.)*

c. GC.
Grave breaches to which the preceding Article relates shall be those involving any of the following acts, if committed against persons or property protected by the present Convention: willful killing, torture or inhuman treatment, including biological experiments wilfully causing great suffering or serious injury to body or health, unlawful deportation or transfer or unlawful confinement of a protected person, compelling a protected person to serve in the forces of a hostile Power, or wilfully depriving a protected person of the rights of fair and regular trial prescribed in the present Convention, taking of hostages and extensive destruction and appropriation of property, not justified by military necessity and carried out unlawfully and wantonly. *(GC, art. 147.)*

503. Responsibilities of the Contracting Parties

No High Contracting Party shall be allowed to absolve itself or any other High Contracting Party of any liability incurred by itself or by another High Contracting Party in respect of breaches referred to in the preceding Article. *(GWS, art. 51; GWS Sea, art. 52; GPW, art. 131; GC, art. 148.)*

504. Other Types of War Crimes

In addition to the "grave breaches" of the Geneva Conventions of 1949, the following acts are representative of violations of the law of war ("war crimes"):

a. Making use of poisoned or otherwise forbidden arms or ammunition.
b. Treacherous request for quarter.
c. Maltreatment of dead bodies.
d. Firing on localities which are undefended and without military significance.
e. Abuse of or firing on the flag of truce.
f. Misuse of the Red Cross emblem.
g. Use of civilian clothing by troops to conceal their military character during battle.
h. Improper use of privileged buildings for military purposes.
i. Poisoning of wells or streams.

j. Pillage or purposeless destruction.
k. Compelling prisoners of war to perform prohibited labor.
l. Killing without trial spies or other persons who have committed hostile acts.
m. Compelling civilians to perform prohibited labor.
n. Violation of surrender terms.

Section IV. DEFENSES NOT AVAILABLE

509. Defense of Superior Orders

a. The fact that the law of war has been violated pursuant to an order of a superior authority, whether military or civil, does not deprive the act in question of its character of a war crime, nor does it constitute a defense in the trial of an accused individual, unless he did not know and could not reasonably have been expected to know that the act ordered was unlawful. In all cases where the order is held not to constitute a defense to an allegation of war crime, the fact that the individual was acting pursuant to orders maybe considered in mitigation of punishment.

b. In considering the question whether a superior order constitutes a valid defense, the court shall take into consideration the fact that obedience to lawful military orders is the duty of every member of the armed forces; that the latter cannot be expected, in conditions of war discipline, to weigh scrupulously the legal merits of the orders received; that certain rules of warfare may be controversial; or that an act otherwise amounting to a war crime may be done in obedience to orders conceived as a measure of reprisal. At the same time it must be borne in mind that members of the armed forces are bound to obey only lawful orders (e. g., *UCMJ, Art. 92).*

510. Government Officials

The fact that a person who committed an act which constitutes a war crime acted as the head of a State or as a responsible government official does not relieve him from responsibility for his act.