Governing Lethal Behavior in Autonomous Robots

Ronald C. Arkin

CRC Press

CHAPTER 3

Human Failings in the Battlefield

THE TREND IS CLEAR: Warfare will continue and autonomous robots will ultimately be deployed in its conduct. Given this, questions then arise regarding if and how these systems can conform as well or better than our soldiers with respect to adherence to the existing Laws of War. This book 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, 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: That is, they do not need to protect themselves in cases of low certainty of target identification. Autonomous armed robotic vehicles 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]. 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 phenomenon 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. These data can arise from multiple remote sensors and intelligence (including human) sources, as part of the Army's network-centric warfare concept [McLoughlin 06] and the concurrent development of the Global Information Grid [DARPA 07]. "Military systems (including weapons) now on the horizon will be too fast, too small, too numerous and will create an environment too complex for humans to direct" [Adams 02].

6. When working in a team of combined human soldiers and autonomous systems as organic assets, 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.

Aside from these ethical considerations, autonomous robotic systems offer numerous other potential operational benefits to the military: faster, cheaper, better mission accomplishment; longer range, greater persistence, longer endurance, higher precision; faster target engagement; and immunity to chemical and biological weapons among others [Guetlein 05]. All of these can enhance mission effectiveness and serve as drivers for the ongoing deployment of these systems. But this book focuses on enhancing ethical benefits by using these systems, ideally without eroding mission performance when compared to human warfighters.

It is not my belief that an autonomous 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. "Armies, armed groups, political and religious movements have been killing civilians since time immemorial" [Slim 08]. Battlefield atrocities* are as old as warfare. "Atrocity ... is the most repulsive aspect of war, and that which resides within man and permits him to perform these acts is the most repulsive aspect of mankind" [Grossman 95].

Man's propensity to wage war has gone unabated for as long as history has been recorded. One could argue that man's greatest failing is being on the battlefield in the first place. Immanuel Kant asserted "War requires no motivation, but appears to be ingrained in human nature and is even valued as something noble" [Kant 85]. Even Albert Einstein, who remained a pacifist well into his fifties, eventually acknowledged "as long as there will be man, there will be war" [Isaacson 07]. Sigmund Freud was even more to the point: "There is no likelihood of our being able to suppress humanity's aggressive tendencies" [Isaacson 07]. In this book, however, we are concerned for the large part with the shortcomings humanity exhibits during the conduct of war (*Jus in Bello*) as opposed to what brought us there in the first place (*Jus ad Bellum*).

"The emotional strain of warfare and combat cannot be quantified" [Bourke 99], but at least there has recently been a serious attempt to gather data on that subject. A recent report from the Surgeon General's Office [Surgeon General 06] assessing the battlefield ethics and mental health 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 noncombatants 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.

* *Atrocity* here is defined as the killing of a noncombatant: either a civilian or a former combatant who has attained *hors de combat* status by virtue of surrender or wound.

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 that 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.

This formal study, although at the very least disconcerting, is by no means the first report of battlefield atrocities. "Atrocious behavior was a feature of combat in the two world wars, as well as in Vietnam" [Bourke 99]. One sociological study of fighting in Vietnam pointed out that, for all men in heavy combat, one-third of men in moderate combat and 8% in light combat had seen atrocities or committed or abetted noncombatant murder [Strayer and Ellenhorn 75]. These numbers are staggering.

Possible explanations for the persistence of war crimes by combat troops are discussed elsewhere [Bill 00, Parks 76, Parks 76a, Danyluk 00, Slim 08]. These include the following:

- 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. This lack of training is not simply in being a good soldier, but also in understanding the Laws of War.
- No clearly defined enemy.
- The issuance of unclear orders where the intent of the order may be interpreted incorrectly as unlawful.
- Shortage of personnel has also been associated in producing stress on combatants that can lead to violations.
- Youth and immaturity of troops.
- An overpowering sense of frustration.
- Pleasure from the power of killing.
- External pressure—for example, for a need to produce a high body count of the enemy.

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

Bourke points out that modern warfare enables violent acts in ways unlike before. Now, "Combatants were able to maintain an emotional distance from their victims largely through the application of ... technology" [Bourke 99]. This portends ill for the reduction of atrocities by soldiers. We now have bombs being dropped in Afghanistan and Iraq by UAV operators from almost halfway around the world in Nevada [CNN 08]. This use of technology enables a form of "numbed killing." She further notes that there is now a "technological imperative" to make full use of the new equipment provided. Although technological warfare has reduced the overall number of soldiers required to wage war, the price is that technology, while increasing the ability to kill, decreases "the awareness that dead

human beings were the end product." When killing at a maximum range, one can pretend they are not killing human beings, and thus experience no regret [Grossman 95]. This physical distance detaches the warfighter from the consequences of the use of their weaponry.

The psychological consequences on our servicemen and women in Afghanistan and Iraq have reached record levels. In 2007 alone, 115 soldiers committed suicide, up from 102 the previous year; 24% of the suicides were those on their first deployment, and 43% were those who had returned from deployment. The suicide rates of active duty soldiers as of August 2008 "were on pace to surpass both last year's numbers and the rate of suicide in the general U.S. population for the first time since the Vietnam war, according to U.S. Army officials" [Mount 08]. A statistically significant relationship has been established between the suicide attempts and the number of days spent deployed in Iraq or Afghanistan. To make matters worse, this is coupled with "a growing number of troops diagnosed with post traumatic stress disorder" [Sevastopulo 08].

These psychiatric casualties are quite significant and common [Grossman 95]: In World War II alone more than 800,000 men were classified unfit due to psychiatric reasons, but an additional 504,000 (approximately fifty divisions) were subsequently rendered unfit as a result of psychiatric collapse after induction. In the 1973 Arab-Israeli war, one-third of the Israeli casualties were psychiatric in origin, twice the number of dead troops. One WWII study showed that after 60 days of continuous combat, 98% of all surviving troops suffered psychiatric trauma of some sort [Swank and Marchand 46]. These long-term exposures to combat are a recent trend in battle, emerging in the twentieth century. The psychiatric damage can result in many forms: battlefield fatigue, conversion hysteria, confusional states, anxiety states, obsession and compulsive states, and character disorders [Grossman 95]. The overall effect on the ability to wage war is obvious, let alone the damage to a nation's surviving citizens.

Creating true warfighters in the first place is a daunting challenge. "No matter how thorough the training, it still failed to enable most combatants to fight" [Bourke 99]. In World War II most men simply did not kill. In one U.S. Army interview of 400 men, only 15% of the men had actually fired at enemy positions (at least once) during an engagement despite the fact that 80% had the opportunity to do so [Marshall 47]. There was no observed correlation between the experience, terrain, nature of the enemy, or accuracy of enemy fire on this percentage.

This applied to both land and air forces. One study of the Korean War indicated that 50% of F-86 pilots never fired their guns and only 10% of those had actually hit a target [Sparks and Neiss 56]. During World War II, most fighter pilots never even tried to shoot anyone down, let alone succeeding. Less than 1% of the pilots accounted for 30–40% of all downed enemy aircraft [Grossman 95].

One conclusion of this is that human soldiers, although not cowardly, lacked an "offensive spirit." One possible reason for this lack of aggressiveness centers on the use of long distance weapons making battlefields "lonely" and the feeling that the enemy was not real but a phantom. This dehumanization of the enemy also quells guilt in killing [Bourke 99].

The soldiers in the field are not alone in their complicity. "Atrocities are the dark secret of military culture" [Danyluk 00]. "Servicemen of all ranks were unperturbed by most of these acts of lawless killing" [Bourke 99]. In Vietnam, combat commanders viewed the Laws of War as "unnecessary" and "unrealistic" restraining devices that would decrease the opportunity for victory [Parks 76]. A lawyer, defending one General's decision not to initiate a court martial for suspected war crimes violations, stated "It's a little like the Ten Commandments—they're there, but no one pays attention to them" [Hersh 71].

Nonetheless our military aspires to higher ethical performance. General Douglas MacArthur stated:

> The soldier, be he friend or foe, is charged with the protection of the weak and unarmed. It is the very essence and reason for his being. When he violates this sacred trust, he not only profanes the cult, but threatens the very fabric of international society. [Hay 76]

In addition the impact of atrocities on public opinion, as clearly evidenced by the My Lai incident in the Vietnam War, and the consequent effect on troop morale are secondary reasons to ensure that events like these are prevented.

Civilians are unfortunately killed during war by other humans for manifold reasons [Slim 08]:

- Genocidal thinking—ethnic or racial cleansing of populations
- Dualistic thinking—separating the good from the bad
- Power dominance and subjugation—power lust and to exert force

- Revenge—emotional striking back for perceived wrongs
- Punishment and forced compliance—to shape the behavior of civilian populations
- Utility—it furthers the war strategically
- Asymmetrical necessity—tactical killing of civilians due to an inferior military position
- Profit—mercenary and looting activity
- Eradicating potential—preemptive removal of civilians that may become warfighters in the future
- Recklessness—shooting anything that moves, or other forms of indiscriminate killing
- Reluctant killing—through human error or accident, collateral damage
- Collective and sacrificial thinking—killing of groups rather than individuals, they must be sacrificed for a greater good

These forms of thinking are alien to current artificial intelligence efforts and likely are to remain so. Armed autonomous systems need not nor should be equipped with any of these forms of unacceptable human rationalization or action.

A primary conclusion is that it seems unrealistic to expect normal human beings by their very nature to adhere to the Laws of Warfare when confronted with the horror of the battlefield, even when trained. As a Marine Corps Reserves Captain commented: "If wars cannot be prevented, steps can be taken to ensure that they are at least fought in as ethical a manner as possible" [Danyluk 00]. One could argue that battlefield atrocities, if left unchecked, may become progressively worse, with the progression of standoff weapons and increasing use of technology. Something must be done to restrain the technology itself, above and beyond the human limits of the warfighters themselves. This is the rationale behind the approach embodied in this book.

Related Philosop Thought

We now turn to several philosophe specifically considered the military's mous robotic agents. Many of them are voca of autonomous battlefield robots. Some ac will ultimately be deployed despite their re calling for an outright ban on the technolo

Interestingly the arguments against au millennia. The crossbow was banned by P against Christians, due to its immoral po enabled killing at a distance [RUSI 08]. Mo struggled into widespread use.

For autonomous lethal robots, we must *autonomy*, as it becomes ambiguous when It is not used here in the strictly philosop the autonomous agent has free will. Here self-directed, and in specific regard to l apropos: "the ability to 'pull the trigger'— out human initiation nor confirmation, b attack command" [Foss 08]. This is restric dier is restricted: the robot soldier must be and any lethal action must be conducted At the highest level, a human is still in the

al striking back for perceived wrongs

orced compliance—to shape the behavior of civil-

the war strategically

essity—tactical killing of civilians due to an infe-
on

and looting activity

tial—preemptive removal of civilians that may
s in the future

oting anything that moves, or other forms of
ing

hrough human error or accident, collateral damage

rificial thinking—killing of groups rather than
ust be sacrificed for a greater good

are alien to current artificial intelligence efforts
n so. Armed autonomous systems need not nor
any of these forms of unacceptable human ratio-

n is that it seems unrealistic to expect normal
ery nature to adhere to the Laws of Warfare when
rror of the battlefield, even when trained. As a
Captain commented: "If wars cannot be prevented,
ure that they are at least fought in as ethical a man-
k 00]. One could argue that battlefield atrocities,
ecome progressively worse, with the progression
increasing use of technology. Something must be
nology itself, above and beyond the human limits
selves. This is the rationale behind the approach

CHAPTER 4

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. Many of them are vocal opponents of the deployment of autonomous battlefield robots. Some acknowledge that these systems will ultimately be deployed despite their reservations, whereas others are calling for an outright ban on the technology.

Interestingly the arguments against automated weaponry date back millennia. The crossbow was banned by Pope Innocent II in 1139 for use against Christians, due to its immoral point-and-click interface, which enabled killing at a distance [RUSI 08]. Most new weapons have similarly struggled into widespread use.

For autonomous lethal robots, we must be clear in our use of the term *autonomy*, as it becomes ambiguous when we cross intellectual disciplines. It is not used here in the strictly philosophical sense, which implies that the autonomous agent has free will. Here we refer to autonomy as being self-directed, and in specific regard to lethality Foss' definition seems apropos: "the ability to 'pull the trigger'—to attack a selected target without human initiation nor confirmation, both in case of target choice or attack command" [Foss 08]. This is restricted only in the same sense a soldier is restricted: the robot soldier must be given a mission to accomplish, and any lethal action must be conducted only in support of that mission. At the highest level, a human is still in the loop so to speak—commanders

must define the mission for the autonomous agent whether it be a human soldier or a robot. The warfighter, robot or human, must then abide by the Rules of Engagement and Laws of War as prescribed from their training or encoding. Autonomy in this sense is limited when compared to a philosopher's point of view.

In a contrarian position regarding the use of battlefield robots, Sparrow 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 [Sparrow 06]. His position is based upon deontological (rights-based) and consequentialist (outcome-based) ethical arguments. He asserts 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, due to the inability to directly control an autonomous robot. Nonetheless, due to the increasing tempo of warfare, he shares my opinion that the eventual deployment of systems with ever increasing autonomy is inevitable. Although I agree that it is necessary that responsibility for the use of these systems must be made clear, I do not agree that it is infeasible to do so. As described in Chapter 2, several existing weapons systems are in use that already deploy lethal force autonomously to some degree, and they (with the exception of antipersonnel 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 direct encoding 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 points out several clear challenges to the roboticist attempting to create a moral sense for a battlefield robot [Sparrow 07]:

- "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 more 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.
  - 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, this problem is indeed solvable. This is further addressed in subsequent chapters.

Asaro similarly argues from a position of loss of attribution of responsibility, but does broach the subject of robots possessing "moral intelligence" [Asaro 06]. His definition of a moral agent seems applicable, where the agent adheres to a system of ethics, which it employs in choosing the actions that it either takes or refrains from taking. He also considers legal responsibility, which he states will compel roboticists to build ethical systems in the future. He notes, similar to what is proposed here, that if an existing set of ethical 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 that we strive to create such an ethical architectural component for unmanned autonomous systems, where that "particular morality" is derived from international conventions.

Regarding *Jus in Bello,* Asaro reminds us that if an autonomous system is potentially capable of reducing collateral damage over previously existing methods of waging war, there is an argument that it is morally required, i.e., a responsibility, to use them [Asaro 07]. The Human Rights Watch group, for example, has stated that only precision-guided bombs should be used in civilian areas [Human Rights Watch 03]. By extension,

if autonomous battlefield robots could reduce civilian casualties over those occasioned by conventional forces, we would be derelict in not using them. Simply stated, at least in some people's view, that if the goals of the research outlined in this book are achieved, i.e., to produce warfighting robots that are more ethical in the battlefield than are human soldiers, a moral imperative exists to deploy such autonomous robotic systems capable of lethal force.

One of the earliest arguments encountered based on 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. Chapter 10 discusses specific issues regarding order refusal overrides by human commanders. Although Perri 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 book 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, 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 [Sullins 06]. 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 book. [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 (nonconscious) 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 book, the robot is off the hook regarding responsibility. We will need to look toward humans for culpability for any ethical errors they make 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. Asaro notes 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* [Asaro 07]. He cites the 1991 Persian Gulf War, the 1999 war in Kosovo, and the 2003 invasion of Iraq as instances where technology made it easier for a nation's leaders and citizens to decide to undertake and support a new war effort. One can argue however, and Asaro does, 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. A primary goal of military research is to provide technological tactical superiority over an opposing force. Thus the argument degenerates to the relinquishing of all military-related research, something that is not likely to happen. As autonomous robotic systems are not envisioned to pose threats similar to those associated with weapons of mass destruction (nuclear, biological, and chemical), it appears unlikely that associated research will be restrained in a similar manner by international convention. A potential arms race could possibly ensue, but again this is a problem for any form of military technology that provides an asymmetric advantage, not simply robotic.

Other *Jus ad Bellum* counterarguments could involve the resulting human-robot battlefield asymmetry as instead having a deterrent effect regarding entry into conflict by the state not in possession of the technology, which now might be more likely to sue for a negotiated diplomatic settlement. 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*. Quite different imagery, one could imagine, as compared to the relatively antiseptic standoff precision high-altitude bombings often seen in U.S. media outlets.

Armstrong is concerned with the impact on the "hearts and minds" of the people in conflict and postconflict zones when and if autonomous robots are deployed [Armstrong 08]. He recalls numerous instances of positive human contact that have helped in reconciling the differences between different cultures, where the presence of robotic technology instead could create a vacuum. In contrast, however, we must note not only the good but also the poor performance of some of our contractors and soldiers in similar circumstances, who have certainly done damage to this cooperative spirit. In any case, a theme that will recur throughout this book is that robots of this sort will not be used in isolation, but rather as organic assets working alongside troops, and not simply replacing them in toto. Human-to-human contact opportunities will persist, just as they have, for example, with the use of canine assets operating side-by-side with soldiers.

Sharkey has been one of the most vocal opponents of autonomous lethal robots, going so far as to calling himself a Cassandra [Sharkey 07, Sharkey 08]. His concerns are manifold: it simply cannot be done correctly because of fundamental limits of artificial intelligence (AI) regarding reliability and discrimination; an echoing of the responsibility concerns voiced by Sparrow and others; the potential for risk-free warfare; and even the cynical point of view that the military will co-opt research such as described in this book "to allay opposition to the premature use of autonomous weapons." Much of his argumentation involves pathos (i.e., it is fear-based), and little logical or formal support is provided for his arguments on AI's limits. Simply because he "has no idea how this could be made to work reliably" does not mean it cannot. The issues surrounding risk-free warfare are addressed below. My personal experience with the integrity of the military allays my concerns regarding co-opting. Besides, the fielding of these systems is likely to proceed independently of whatever efforts are undertaken in regard to ethically embedding an "artificial conscience," no doubt by using more conventional approaches to manage the legality

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

of this new class of weapons. Sharkey and I both agree, however, that the time has come to discuss these issues on an international scale, to determine what and if any limits should be applied to battlefield use of lethal autonomous systems.

Borenstein takes a more reasoned stance, revisiting some of the concerns already raised [Borenstein 08]. To those he adds the unforeseen problems associated with software glitches, some of which have already resulted in significant deaths. He cites software problems surrounding the death of 28 Americans when a missile defense system failed [GAO 92] and a South African automated antiaircraft system that went out of control resulting in the deaths of nine soldiers [Hosken et al. 07]. He also notes that humans have situational and instinctual knowledge to rely on that will be difficult to encode in a robotic system, well above and beyond the Laws of War. Although this is currently true, it may not be a limit of the future, but in any case it should not serve as a deterrent to restrain the use of force by autonomous systems provided with existing well-defined laws, as these systems are seemingly inevitably being deployed. Other concerns (e.g., technological vulnerability such as hacking) are more easily dismissed with the ongoing major efforts by the DOD in cybersecurity. Although Borenstein remains skeptical, he does cede that "If advances in AI do continue to move forward, reaching close to duplicating the human brain, some of the fears relating to AWS [Autonomous Weapons Systems] might conceivably lessen" [Borenstein 08]. Nonetheless, his *Jus ad Bellum* concerns regarding "escalation and removing potential deterrents to war" persist.

Sparrow has recently commented on the requirement that UV systems be designed to be ethical from the onset, focusing on the responsibility of the designer to ensure that these systems are built to be safe and to incorporate the Laws of War [Sparrow 08]. One key aspect is his focus on the design of an interface for operators that enforces morality, building ethics into the system directly. "The interface for an [Unmanned System] should facilitate killing where it is justified and frustrate it where it is not," a challenge, as he puts it, that is yet to be met. We share this concern and seemingly agree on the value of embedding ethics into both the robotic system itself and its operator interface.

Another often heard argument against the use of autonomous weapons is that they will be incapable of exhibiting mercy, compassion, and humanity [Davis 08]. Although substantial progress is being made in artificial intelligence on the use of emotions in robotic systems (e.g., [Fellous and Arbib 05]), and indeed guilt and remorse are recommended for

implementation within the architecture presented in subsequent chapters, no current provision is made for these emotions at this time. The rationale is not because it is more challenging than other secondary emotions, but rather that humanity is legislated into the Laws of War, and as such if they are followed, the robot will exercise restraint consistent with societal norms. This may be inadequate to some, but the reduction of the inhumanity exhibited by a significant percentage of soldiers [Surgeon General 06] is believed to offset this loss and can potentially result in a fighting force that is more humane overall than an all-human one.

Potential proliferation of the underlying technology has also been expressed as a concern. Rear Admiral Chris Parry of the U.K. Royal Navy broached this subject at a recent workshop [Parry 08]. The ease with which unmanned drones can be made from hobby aircraft kits coupled with GPS and cell phone technology is just one example that would enable terrorists to easily manufacture buzz-bomb-type UAVs for use against events such as the upcoming Olympics in London. Frightening prospects indeed. It was reported that Hezbollah launched two attack UAVs against Israel on August 13, 2006, with at least one apparently armed with 30 kg of explosive that was recovered at the wreckage site [Eshel 06]. They were intercepted by the Israeli Air Force before they reached their target. Clearly, you need not be a major international power to take advantage of the underlying technology. These worrisome aspects of proliferation need ongoing attention.

One argument voiced by military personnel regarding the introduction of ethical autonomous robots into the battlefield is the potential for a deleterious effect on squad cohesion. This term refers to the "Band of Brothers" attitude formed by a small group of men in combat, who come to rely on and protect each other. If a robot that is capable of objectively monitoring the moral performance of team members is injected into the unit, it may seriously impede the effectiveness of the team due to a fracturing of trust. The concept of even "fragging" the robot has been mentioned, where it would be deliberately destroyed by squad members to prevent infractions from being reported. The counterargument for this possible effect may lie within the performance of the robot itself: if it is willing to go out in advance of my men, if it is willing to take a bullet for me, if it can watch my back better than a fellow human soldier could, then the omnipresent ethical monitoring might be a small price to pay in favor of my enhanced survival. Attention would need to be focused on how to establish this level of human-robot trust, but through experience and training it should be feasible to establish a meaningful bond between man and machine.

For example, consider one robot's story used for removing improvised explosive devices in Iraq:

> After several successful missions, the Packbot ... was destroyed. The operator brought it back to the makers and asked for it to be rebuilt. He didn't want a new one, he wanted it fixed. It was a good robot and they'd been through a lot together. [Bains 07]

The United States Navy is examining the legal ramifications of the deployment of autonomous lethal systems in the battlefield [Canninget al. 04], observing that a legal review is required of any new weapons system prior to its acquisition to ensure that it complies with the LOW and related treaties. To pass this review, it must demonstrate that it neither acts indiscriminately nor causes 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, unless 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 or a lack of recognition of neighboring civilian objects, is legal in itself (i.e., ensuring that proportionality is exercised against a military objective). Canning 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, Canning 08]. Concerns arise from current limits on the ability to discriminate combatants from noncombatants on the battlefield. 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 in this research area, and in many ways it already has begun, e.g., by using gait recognition and other patterns of activity to identify suspicious persons. These 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 in the battlefield. This is an acceptable strategy by virtue of the robot's not needing to defend itself as a soldier would, perhaps even 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 ideas, design, and results presented in this book constitute only 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 thus can and should 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*, 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]. Walzer 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; they are not the ends or the means to the ends. He argues for a stronger

* The Principle (or Doctrine) 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" [Norman 95, Wells 96]. 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).

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 casualties), 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 (e.g., rifle versus grenade), targeting accuracy (standoff distances) in the presence of civilian populations, or other similar criteria. Walzer 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]

Anderson, 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 [Anderson, K 07]. 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 to be considered [Kolodner 93]. 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.

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]. 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's "Three Laws of Robotics"* [Asimov 50] (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 straw man to bootstrap the ethical debate and as such serve no useful practical purpose beyond their fictional roots. Anderson 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" [Anderson 07b]. With all due respect, I must concur.

* See http://en.wikipedia.org/wiki/Three_Laws_of_Robotics for a summary discussion of all four laws.

Other factors can further define the overall situation such as intention (plans from the deliberative component of the architecture) and internal motivations (endogenous factors such as fuel levels, affective state, etc.).

A new behavioral coordination function, **C**, is now defined such that the overall robotic response ρ is determined by:

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

or alternatively:

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

where

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

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 ρ.

6.2 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 ρ_{ij}, where $\rho_{ij} \in \boldsymbol{P}$ (read as capital rho), and $\boldsymbol{P}$ denotes the set of all possible overt responses. $\rho_{lethal\text{-}ij}$ 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 $\rho_{lethal\text{-}ij}$. A subset $\boldsymbol{P}_{l\text{-}ethical}$ of $\boldsymbol{P}_{lethal}$ can be considered the set of *ethical* lethal behaviors if for all discernible **S**, any $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ produced bv $\beta_{lethal\text{-}i}$ 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 (**S**) potentially encountered by the autonomous agent in a given mission context. 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 executed 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 $\mathbf{S}_j$; or
2. a positive behavioral constraint (an obligation) that requires a behavior $\beta_{lethal\text{-}i}$ to produce $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ in a given perceptual situational context $\mathbf{S}_j$.

* The Closed World Assumption, from artificial intelligence, presumes that whatever is not currently known to be true is false.

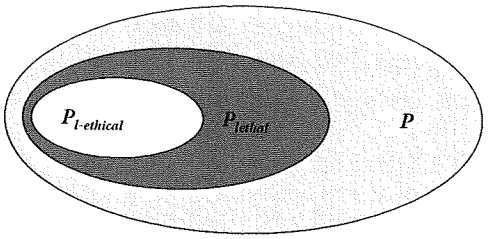

FIGURE 6.1 Behavioral action space ($P_{l\text{-}ethical} \subseteq P_{lethal} \subseteq P$).

Discussion of the specific representational choices for these constraints C is deferred until Chapter 10.

Now consider Figure 6.1, where P denotes the set of all possible overt responses ρ_{ij} (situated actions) generated by the set of all active behaviors $\mathbf{B}$ for all discernible situational contexts $\mathbf{S}$ for a given robot; P_{lethal} is a subset of P which includes all actions involving lethality, and $P_{l\text{-}ethical}$ is the subset of P_{lethal} representing all ethical lethal actions that the autonomous robot can undertake in all given situations $\mathbf{S}$. $P_{l\text{-}ethical}$ is determined by C being applied to 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.

$P_{lethal} - P_{l\text{-}ethical}$ is denoted as $P_{l\text{-}unethical}$, where $P_{l\text{-}unethical}$ is the set of all individual $\rho_{l\text{-}unethical\text{-}ij}$ unethical lethal responses for a given $\mathbf{B}_{\text{lethal-i}}$ in a given situation $\mathbf{S}_j$. These unethical responses must be avoided in the architectural design through the application of C onto P_{lethal}. $P - P_{l\text{-}unethical}$ forms the set of all permissible overt responses $P_{permissible}$, which may be lethal or not. Figure 6.2 illustrates these relationships.

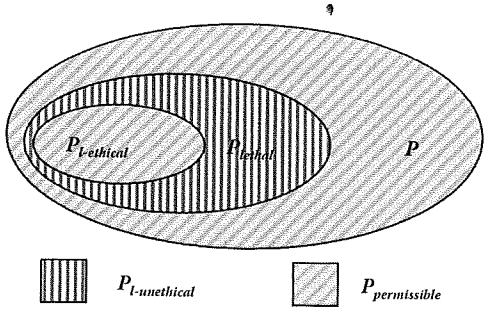

FIGURE 6.2 Unethical and permissible actions regarding the intentional use of lethality (compare to Figure 6.1).

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

1. *Ethical Situation Requirement:* Ensure that only situations $\mathbf{S}_j$ that are governed (spanned) by C can result in $\rho_{lethal\text{-}ij}$ (a lethal action for that situation). Lethality cannot result in any other situations.
2. *Ethical Response Requirement (with respect to lethality):* Ensure that only permissible actions $\rho_{ij} \in P_{permissible}$, 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).
3. *Unethical Response Prohibition:* Ensure that any response $\rho_{\text{l-unethical-ij}} \in P_{l\text{-}unethical}$, is either:
   a. mapped onto the null action $\boldsymbol{\varnothing}$ (i.e., it is inhibited from occurring if generated by the original controller);
   b. transformed into an ethically acceptable action by overwriting the generating unethical response $\rho_{l\text{-}unethical\text{-}ij}$, perhaps by a stereotypical nonlethal action or maneuver, or by simply eliminating the lethal component associated with it; or
   c. precluded from ever being generated by the controller in the first place by suitable architectural design through the direct incorporation of C into the design of $\mathbf{B}$.
4. *Obligated Lethality Requirement:* In order for a lethal response $\rho_{lethal\text{-}ij}$ to result, there must exist at least one constraint c_k derived from the ROE that obligates the use of lethality in situation $\mathbf{S}_j$.
5. 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 the combatant/noncombatant discrimination requirements of *Jus in Bello.*

We will see that these conditions result in several alternative architectural choices for the design and implementation of an ethical lethal autonomous system (see Chapter 10 for an expanded discussion of each of these approaches):

1. *Ethical Governor:* which suppresses, restricts, or transforms any lethal behavior $\rho_{lethal\text{-}ij}$ (ethical or unethical) produced by the existing architecture so that it must fall within $P_{permissible}$ after it is initially

generated by the architecture (post facto). This means if $\rho_{l\text{-}unethical\text{-}ij}$ 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; that is, it is transformed to $\rho_{permissible\text{-}ij}$.

2. *Ethical Behavioral Control*: which constrains all active behaviors (β_1, β_2, ... β_m) in **B** to yield **R** with each vector component $\mathbf{r}_i \in \boldsymbol{P}_{permissible}$ set as determined by C; that is, only lethal ethical behavior is produced by each individual active behavior that involves lethality in the first place.

3. *Ethical Adaptor:* if a resulting executed lethal behavior is post facto determined to have been unethical, that is, $\rho_{ij} \in \boldsymbol{P}_{l\text{-}unethical}$, then the system must 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), for example, via an after-action reflective review or through the application of an artificial affective function (e.g., guilt, remorse, or grief).

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 only 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 a deliberative/reactive hybrid autonomous robot architecture such as AuRA [Arkin and Balch 97], similar to one under development 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, the **Responsibility Advisor** (Chapter 10), should make clear and explicit as best as possible, just where *responsibility* vests, if: (1) an unethical action within the space $\boldsymbol{P}_{l\text{-}unethical}$ 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 inadvertent or deliberate

representational deficiency in the constraint set C or in the system's application outside of an appropriate mission context either by the operator or from 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 their deployment of a lethal autonomous system. The robot architecture must be capable to some degree of providing suitable explanations for its actions regarding lethality (including refusals to act).

Chapter 10 forwards architectural specifications for handling all these design alternatives above, and Chapter 12 presents some prototype implementation results driven from those specifications. 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 book, and it is left for future work.

described in Chapter 11. As appropriate, provision is made in the overall architecture for the underlying behaviors to have access to the global constraint set C as needed (Figure 10.1). This may be especially important for the choice of short-term memory representations regarding the ROE.

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.

10.3 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, should difficulties arise.

The Ethical Adaptor operates at two primary levels:

1. *After-action reflection,* where reflective consideration and critiquing of the performance of the lethal robotic system, triggered either by a human specialized in such assessments or by the system's post-mission cumulative internal affective state (e.g., guilt or remorse), provides 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.

10.3.1 After-Action Reflection

This ethical adaptor component involves introspection 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 making a legally correct ethical assessment regarding the appropriateness of the autonomous agent's operation in

the given situation. The greatest benefit of this procedure will likely be derived during the robot's training exercises, 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 states that for human soldiers "experience is the best trainer. The draft scenarios could structure experiences challenging the memorized RAMP rules to the real world" [Martins 94]. In addition, if the autonomous agent has imposed affective restrictions upon itself during the mission, after-action reflection upon these violated expectations must be performed to ensure that these events do not recur.

This essentially is a form of one-shot learning (no pun intended) involving constraint specialization (a form of restriction). The revision methods will operate over externalized variables of the underlying behaviors, using methods similar to those employed in a Phase I project recently performed for the Navy jointly with Mobile Intelligence Inc., entitled *Affect Influenced Control of Unmanned Vehicle Systems* [OSD 06]. For the ethical architecture, 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 parameters (use less destructive force)
5. Weapon firing patterns (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, whereas Items 4–6 are concerned with proportionality and the Principle of Double Intention. These values must always be altered in a manner to become more restrictive, as they are being altered

as a result of perceived ethical infractions. Determination of the offending constraints or parameters will, at least initially, require human intervention and guidance, as credit assignment is a well-known problem for artificial intelligence.* 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, not always seen in humans.

10.3.2 Affective Restriction of Behavior

It was observed earlier, that human emotion has been indicted in creating the potential for war crimes, so one might wonder why we are even considering the use of affect at all. What is proposed here is the use of a strict subset of affective components, those that are specifically considered the moral emotions [Haidt 03]. Indeed, 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 design component modeling a subset of these affective components (initially only guilt) is intended to provide an adaptive learning function for the autonomous system architecture should it act in error.

* The credit assignment problem in artificial intelligence refers to how credit or blame is assigned to a particular piece or pieces of knowledge in a large knowledge base or to the component(s) of a complex system responsible for either the success or failure in an attempt to accomplish a task.

Haidt provides a taxonomy of moral emotions [Haidt 03]:

- 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 toward the self (i.e., the autonomous agent), and in particular guilt, which should be produced whenever suspected violations of the ethical constraint set *C* occur or from direct criticism received from human operators or authorities regarding its own 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 adopt for use in the ethical governor. In our design, guilt should only result from unintentional effects of the robotic agent, but nonetheless its presence should alter the future behavior of the system so as to eliminate or at least minimize the likelihood of recurrence 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, Arkin et al. 03]. The design and implementation of the ethical architecture draws upon this experience. It is intended initially to solely manage the single affective variable of guilt (V_{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 self-monitoring processes that the system may be able to maintain on its own (again, assuming autonomous perceptual capabilities can achieve that level of performance), e.g., battle damage assessment of noncombatant casualties and damage to civilian property, among others.

Should any of these perceived ethical violations occur, the affective value of V_{guilt} will increase monotonically until the after action review is undertaken. If these cumulative affective values (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:

$$\text{IF } V_{guilt} > \text{Max}_{guilt} \quad \text{THEN } \boldsymbol{P}_{l\text{-}ethical} = \boldsymbol{\emptyset}$$

where V_{guilt} represents the current scalar value of the affective state of guilt, and Max_{guilt} is a threshold constant. This denial-of-lethality step is irreversible for as long as the system is in the field, and once triggered, it is independent of any future value for V_{guilt} until an after-action review. It may be possible for the operators to override this restriction, if they are willing to undertake that responsibility explicitly and submit to an ultimate external review of such an act (Chapter 12). In any case, the system can continue operating in the field, but only in a nonlethal support capacity if appropriate (e.g., for reconnaissance or surveillance). It is not necessarily required to withdraw from the field, but it can only serve henceforward without any further potential for lethality. 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 instead involves the emotion of shame). The value 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 book. One study provides a social contract ethical framework involving moral values that include guilt, which addresses the problem of work distribution among parties [Cervellati et al. 07]. Another effort developed a dynamic model of guilt for understanding motivation in prejudicial contexts [Amodio et al. 07]. 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 toward 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 considerably 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 architectural approach is the ability to introduce compassion as an emotion at this time, which may be considered by some as a serious deficit in a battlefield robot. While it is less clear how to introduce such a capability, by requiring the autonomous system to abide strictly to the LOW and ROE, we 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.

10.4 RESPONSIBILITY ADVISOR

"If there are recognizable war crimes, there must be recognizable criminals" [Walzer 77]. 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 even for human war criminals, we need to ensure that accountability is built into the ethical architecture of an autonomous system to support such needs.