Governing Lethal Behavior:
Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture
Part 3: Representational and Architectural Considerations[1]

Ronald C. Arkin
Mobile Robot Laboratory
College of Computing
Georgia Institute of Technology

ABSTRACT: This paper, the third in a series, provides representational 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

This article presents ongoing research funded by the Army Research Office on providing an ethical basis for autonomous system deployment in the battlefield, specifically regarding the potential use of lethality. This project entitled "An Ethical Basis for Autonomous System Deployment". It is concerned with two research thrusts addressing the issues of autonomous robots capable of lethality:

1) What is acceptable? Can we understand, define, and shape expectations regarding battlefield robotics? Toward that end, a survey has been conducted to establish opinion on the use of lethality by autonomous systems spanning the public, researchers, policymakers, and military personnel to ascertain the current point-of-view maintained by various demographic groups on this subject.
2) What can be done? We are designing a computational implementation of an ethical code within an existing autonomous robotic system, i.e., an "artificial conscience", that will be able to govern an autonomous system's behavior in a manner consistent with the rules of war.

The results obtained for question (1) are presented in [Moshkina and Arkin 07]. Question (2) is addressed in a series of papers of which this is the third. Part I of this series [Arkin 08a] discusses the motivation and philosophy for the design of such a system, incorporating aspects of the Just War tradition [Walzer 78], which is subscribed to by the United States. It presents the requirements of military necessity, proportional use of force, discrimination, and responsibility attribution, and the need for such accountability in unmanned systems, as the use of autonomous lethality appears to progress irrevocably forward. Part II [Arkin 08b] presents the theoretical formalisms used to help specify the overall design of an ethical architecture that is capable of incorporating the Laws of War (LOW) and Rules of Engagement (ROE) as specified by International Law and the U.S. Military. This paper (Part III) provides a description of the representational basis for the implementation of such a system. A complete compilation of the material presented in this series appears in a lengthy technical report [Arkin 07].

2. Representational Considerations for the Ethical Control of Lethality

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

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 Laws of War (LOW) [US Army 56, 72, Berger et al 04] and Rules of Engagement (ROE) [CLAMO 00] must be maintained, strictly adhering to the rights of noncombatants regarding target discrimination (deontological), while considering similar principles for the assessment of proportionality of force based on military necessity (utilitarian). 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 under development in this research uses an action-based approach, where ethical theory (as encoded in the LOW and ROE) informs the agent what actions to undertake or not to undertake. Action-based methods have the following attributes [Anderson et al. 06]: Consistency – the avoidance of contradictions in the informing theory; Completeness – how to act in any ethical dilemma; Practicality – it should be feasible to execute, and agreement with expert ethicist intuition

None of these appears 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 the agent 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 described later.

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] 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. 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.

Drawing on the set-theoretic descriptions developed in Part II of this series of papers [Arkin 08b]:

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 $\boldsymbol{P_{l\text{-}unethical} = P_{lethal} - P_{l\text{-}ethical}}$, 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 below.

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.

A particular constraint 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 a lethal response $r_{lethal\text{-}ij}$ for a given perceptual situation S_j.
2. a positive behavioral constraint (an obligation) which requires a behavior $\beta_{lethal\text{-}i}$ to produce an ethical lethal response $r_{l\text{-}ethical\text{-}ij}$ 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 obligatory 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 during development 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 [Arkin 08a], its inevitable introduction.

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. We now 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.

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 (1) a robot take only permissible actions and (2) that obligatory actions are indeed performed, subject to ties and conflicts among available actions. 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.

The ethical code $\boldsymbol{C}$ a robot uses in general 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 **L** is represented as $\boldsymbol{\Phi}_{\boldsymbol{C}}^{\mathbf{L}}$. This basically reduces the problem for our ethical governor (the architectural means by which ethical behavior is enforced) to the need to derive from the LOW and ROE a suitable $\boldsymbol{\Phi}^{\mathbf{L}}_{\mathbf{LOW} \cup \mathbf{ROE}}$, with the leading candidate for **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.

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.

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], for example, implemented an ethical reasoning system called Jeremy that is capable of conducting moral arithmetic. 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 considered inappropriate for our goals.

[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. 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 nor should have the machine derive its ethical rules on its own.

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 autonomous system 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. Another alternative CBR-based approach (the W.D. system) that uses a duty-based system was developed by [Anderson et al. 06] that *does* arrive at ethical conclusions derived from case data. Rules (principles) are derived from cases provided by an expert ethicist who serves as a trainer. These rules are generalized as appropriate. [Andersen et al. 05] developed a similar system, MedEthEx, for use in the medical ethics domain to serve as an advisor.

3. Architectural Considerations

We now move closer towards an implementation of the underlying theory using, as appropriate, the content and format of the representational knowledge described in [Arkin 08b]. 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 requirement that compliance with the Laws of War 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, just how is that enforced upon an underlying behavioral 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 a whole and 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.

3.1 Architectural Requirements

In several respects, the design of an ethical 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. 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]. The onset of autonomy in the battlefield is not a discontinuous event but rather follows a smooth curve, permitting a gradual and subtle introduction of this capability into the battlefield as the technology progresses, a form of technology creep.

[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 1.

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. 1: 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 system than that of a human? Unfortunately, we have already observed in there is plenty of room for improvement [Surgeon General 06, Arkin 08a]. 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 [Walzer 77], 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 the application of lethal force, which becomes 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 thresholds τ should be defined appropriately to only permit lethality in cases where clear confirmation of a discriminated target is available and ideally supported by multiple sources of evidence.

Considering our earlier discussion on forbidden and obligatory actions, 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 the 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 2. 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.

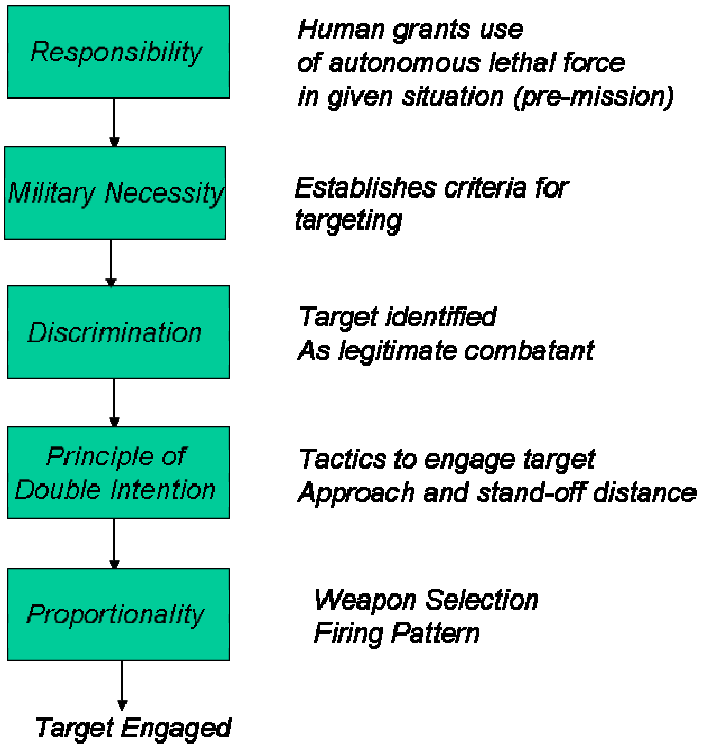

Figure 2: Ethical Architectural Principle and Procedure

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:

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

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:

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

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.

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]

3.2 Architectural Design Options

We turn 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. 3). 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 ($\mathbf{r_{l\text{-}ethical\text{-}ij}}$).
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 the direction of a more restrictive manner. It is based upon both an after-action reflective review of the system's performance and 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 (cf. [Arkin 05]).
4. **Responsibility Advisor**: This component forms a part of the human-robot interaction interface 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 that prevents the autonomous agent from acting.

The preliminary specifications and design for each of these system components is described in more detail in [Arkin 07]. 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 (a 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 are currently made regarding the ultimate solvability of these problems, and as such this should temper any optimism involving the ability to field an ethical autonomous agent capable of lethality in the near term.

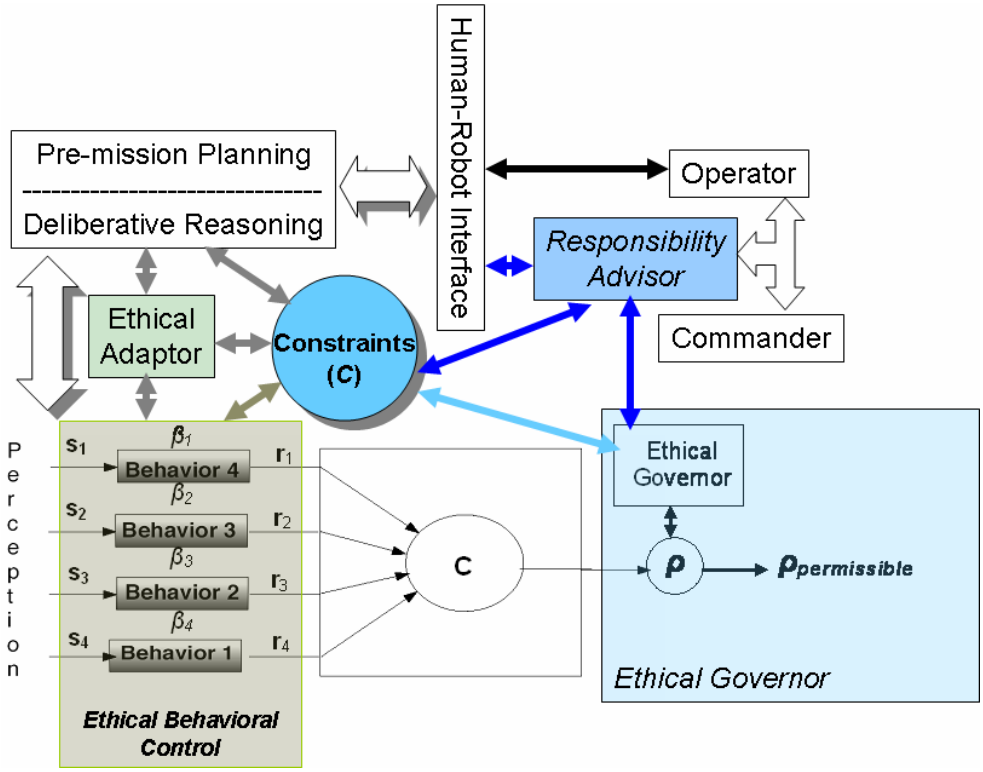

Figure 3: Major Components of an Ethical Autonomous Robot Architecture. The newly developed ethical components are shown in color.

4. 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 additional 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 [Baldor 06]. 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 a wounded 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 [Samsung 07], 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 an unmanned system 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.

Considerably more detail on each of these scenarios can be found in [Arkin 07].

5. Summary, Conclusions, and Future Work

This report has provided the representational requirements, architectural design criteria and recommendations to design and construct an autonomous robotic system architecture capable of the ethical use of lethal force. These first steps toward that goal are 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 provides 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, and should those norms be violated in any way, invoke 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 paper. 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.

6. References

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

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

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

Arkin, R.C., Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture – Part 1: Motivation and Philosophy, to appear in *Proc. HRI 2008*, Amsterdam, NL, 2008a.

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

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

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..

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.

Baldor, L., "Military Declined to Bomb Group of Taliban at Funeral", AP News Story, 9/15/2006.

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.

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.

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

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

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

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

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

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

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

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

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

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.

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.

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., "Lethality and Autonomous Systems: Survey Design and Results", GVU Technical Report GIT-GVU-07-16, College of Computing, Georgia Tech, 2007.

Parks, W.H., "Commentary", 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.

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

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.

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

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 Pres*s, 2005.

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

U.S. Army, Pamphlet 27-161-2, *International Law, Volume II* (23 October 1962)

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

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.