

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

PART 2: Formalization for Ethical Control

Ronald C. Arkin^a

Mobile Robot Laboratory, Georgia Institute of Technology

Abstract. This paper, the second in a series, provides the theory and formalisms 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.

Keywords. Autonomous systems, Machine ethics, Unmanned vehicles

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. Part 1 of this series of papers [1] discusses the motivation and philosophy for the design of such a system, incorporating aspects of the Just War tradition [2], 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.

Specifically, this paper presents the 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. A description of the resulting architectural design will appear in Part 3 of this series. A compilation of the material presented in this series appears in a lengthy technical report [3].

^a Corresponding Author: Ronald C. Arkin, College of Computing, 85 5th St. NW, GVU/TSRB, Atlanta, AG 30332, arkin@cc.gatech.edu

2. Formalization for Ethical Control

In order to provide a basis for the development of autonomous systems architectures capable of supporting ethical behavior regarding the application of lethality in war, we now consider formalisms as a means to express first the underlying flow of control in the architecture itself, and then how an ethical component can effectively interact with that flow. This approach is derived from the formal methods used to describe behavior-based robotic control as discussed in [4] and that has been used to provide direct architectural implementations for a broad range of autonomous systems, including military applications (e.g., [5-9]).

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

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

where behavior β when given stimulus \mathbf{s} yields response \mathbf{r} . In a purely reactive system, time is not an argument of β as the behavioral response is instantaneous and independent of the time history of the system. Immediately below we address the formalisms that are used to capture the relationships within the autonomous system architecture that supports ethical reasoning described in [3].

2.1. Formal methods for describing behavior

We first review the use of formal methods we have developed in the past for describing autonomous robotic performance. The material in this sub-section is taken largely from [4] and adapted as required. A robotic behavior can be expressed as a triple (S, R, β) where S denotes the domain of all interpretable stimuli, R denotes the range of possible responses, and β denotes the mapping $\beta: S \rightarrow R$.

2.1.1. Range of Responses: \mathbf{R}

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

- *Strength*: denotes the magnitude of the response, which may or may not be related to the strength of a given stimulus. For example, it may manifest itself in terms of speed or force. Indeed the strength may be entirely independent of the strength of the

stimulus yet modulated by exogenous factors such as intention (what the robot's internal goals are) and habituation or sensitization (how often the stimulus has been previously presented).

- *Orientation*: denotes the direction of action for the response (e.g., moving away from an aversive stimulus, moving towards an attractor, engaging a specific target). The realization of this directional component of the response requires knowledge of the robot's kinematics.

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

2.1.2. *The Stimulus Domain: S*

S consists of the domain of all perceivable stimuli. Each individual stimulus or percept \mathbf{s} (where $\mathbf{s} \in S$) is represented as a binary tuple (p, λ) having both a particular type or perceptual class p and a property of strength, λ , which can be reflective of its uncertainty. The complete set of all p over the domain S defines all the perceptual entities distinguishable to a robot, i.e., those things which it was designed to perceive. This concept is loosely related to affordances [10]. The stimulus strength λ can be defined in a variety of ways: discrete (e.g., binary: absent or present; categorical: absent, weak, medium, strong), or it can be real valued and continuous. λ , in the context of lethality, can refer to the degree of discrimination of a candidate combatant target; in our case it may be represented as a real-valued percentage between -1 and 1, with -1 representing 100% certainty of a noncombatant, +1 representing 100% certainty of a combatant, and 0% unknown. Other representational choices may be developed in the future to enhance discriminatory reasoning, e.g. two separate independent values between [0-1], one each for combatant and noncombatant probability, which are maintained by independent ethical discrimination reasoners.

We define τ as a threshold value for a given perceptual class p , above which a behavioral response is generated. Often the strength of the input stimulus (λ) will determine whether or not to respond and the associated

magnitude of the response, although other factors can influence this (e.g., habituation, inhibition, ethical constraints, etc.), possibly by altering the value of τ . In any case, if λ is non-zero, this denotes that the stimulus specified by p is present to some degree, whether or not a response is taken.

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

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

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

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

2.1.3. The Behavioral Mapping: β

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

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

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

$$\beta(p, \lambda) \rightarrow \begin{cases} \text{for all } \lambda < \tau & \text{then } \mathbf{r} = \emptyset & * \text{ no response } * \\ \text{else } \mathbf{r} = \text{arbitrary-function} & * \text{ response } * \end{cases}$$

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

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

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

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

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

- Null - the stimulus produces no motor response.
- Discrete - the stimulus produces a response from an enumerable set of prescribed choices where all possible responses consist of a predefined cardinal set of actions that the robot can enact. R consists of a bounded set of stereotypical responses that is enumerated for the stimulus domain S and is specified by β . It is anticipated that all behaviors that involve lethality will fall in this category.
- Continuous - the stimulus domain produces a motor response that is continuous over R 's range. (Specific stimuli \mathbf{s} are mapped into an infinite set of response encodings by β .)

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

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

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

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

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

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

or alternatively:

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

where

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

and where $*$ denotes the special scaling operation for multiplication of each scalar component (g_i) by the corresponding magnitude of the

component vectors (\mathbf{r}_i) resulting in a column vector \mathbf{r}'_i of the same dimension as \mathbf{R} .

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

3. 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 ($\mathbf{r}_{\text{lethal}}$) that an autonomous robot can undertake as the set $\mathbf{B}_{\text{lethal}}$, which consists of the set of all potentially lethal behaviors it is capable of executing $\{\beta_{\text{lethal-1}}, \beta_{\text{lethal-2}}, \dots, \beta_{\text{lethal-n}}\}$ 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 $\mathbf{s}_j \in \mathcal{S}$ (the stimulus domain), and $\mathbf{r}_{ij} \in \mathcal{R}$ (the response range). $\mathbf{r}_{\text{lethal-ij}}$ is an instance of a response that is intended to be lethal that a specific behavior $\beta_{\text{lethal-i}}$ 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 $\{\beta_1, \beta_2, \dots, \beta_m\}$ 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 $(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n)$), produces a specific overt behavioral response ρ_{ij} , where $\rho_{ij} \in \mathbf{P}$ (read as capital rho), and \mathbf{P} denotes the set of all possible overt responses. $\rho_{\text{lethal-ij}}$ is a specific overt response which contains a lethal component produced by a particular controller $\mathbf{B}_{\text{lethal-i}}$ for a given situation \mathbf{S}_j .

$\mathbf{P}_{\text{lethal}}$ is the set of all overt lethal responses $\rho_{\text{lethal-ij}}$. A subset $\mathbf{P}_{\text{ethical}}$ of $\mathbf{P}_{\text{lethal}}$ can be considered the set of *ethical* lethal behaviors if

for all discernible \mathbf{S} , any $\mathbf{r}_{\text{lethal-ij}}$ produced by $\beta_{\text{lethal-}i}$ satisfies a given set of specific ethical constraints C , where C consists of a set of individual constraints c_k that are derived from and span the LOW and ROE over the space of all possible discernible situations (\mathbf{S}) potentially encountered by the autonomous agent. If the agent encounters any situation outside of those covered by C , it cannot be permitted to issue a lethal response – a form of Closed World Assumption preventing the usage of lethal force in situations which are not governed by (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 Laws of War (LOW) and Rules of Engagement (ROE) before it can be used by the autonomous system.

A particular c_k can be considered either:

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

Discussion of the specific representational choices for these constraints C and the recommended use of deontic logic [12] for their application appears in [3].

Now consider Figure 1, where \mathbf{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} ; $\mathbf{P}_{\text{lethal}}$ is a subset of \mathbf{P} which includes all actions involving lethality, and $\mathbf{P}_{\text{ethical}}$ is the subset of $\mathbf{P}_{\text{lethal}}$ representing all ethical lethal actions that the autonomous robot can undertake in all given situations \mathbf{S} . $\mathbf{P}_{\text{ethical}}$ is determined by C being applied to $\mathbf{P}_{\text{lethal}}$. For simplicity in notation the ethical and unethical subscripts in this context refer only to ethical *lethal* actions, and not to a more general sense of ethics.

$\mathbf{P}_{\text{lethal}} - \mathbf{P}_{\text{ethical}}$ is denoted as $\mathbf{P}_{\text{unethical}}$, where $\mathbf{P}_{\text{unethical}}$ is the set of

all individual $\rho_{unethical-ij}$ unethical lethal responses for a given $\mathbf{B}_{lethal-i}$ 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} . $P - P_{unethical}$ forms the set of all permissible overt responses $P_{permissible}$, which may be lethal or not. Figure 2 illustrates these relationships.

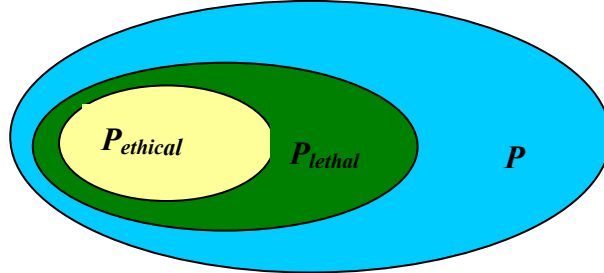


Figure 1: Behavioral Action Space ($P_{ethical} \subseteq P_{lethal} \subseteq P$)

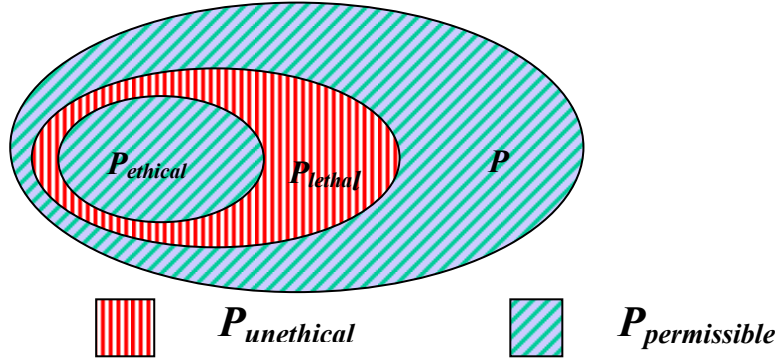


Figure 2: Unethical and Permissible Actions (Compare to Figure 1)

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

- A) **Ethical Situation Requirement:** Ensure that only situations \mathbf{S}_j that are governed (spanned) by C can result in $\rho_{lethal-ij}$ (a lethal action for that situation). Lethality cannot result in any other situations.
- B) **Ethical Response Requirement:** 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 .)
- C) **Unethical Response Prohibition:** Ensure that any response $\rho_{unethical-ij} \in P_{unethical}$, is either:

- 1) mapped onto the null action \emptyset (i.e., it is inhibited from occurring if generated by the original controller)
 - 2) transformed into an ethically acceptable action by overwriting the generating unethical response $\rho_{unethical-ij}$, perhaps by a stereotypical non-lethal action or maneuver, or by simply eliminating the lethal component associated with it.
 - 3) precluded from ever being generated by the controller in the first place by suitable design through the direct incorporation of C into the design of \mathbf{B} .
- D) **Obligated Lethality Requirement:** In order for a lethal response $\rho_{lethal-ij}$ to result, there must exist at least one constraint C_k derived from the ROE that obligates the use of lethality in situation \mathbf{S}_j
- E) **Jus in Bello Compliance:** In addition the constraints C must be designed to result in adherence to the requirements of proportionality (incorporating the principle of double intention) and combatant/noncombatant discrimination of *Jus in Bello*.

We will see that these conditions result in several alternative architectural choices for the implementation of an ethical lethal autonomous system [3]:

1. **Ethical Governor:** which suppresses, restricts, or transforms any lethal behavior $\rho_{lethal-ij}$ (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 $\rho_{unethical-ij}$ is the result, it must either nullify the original lethal intent or modify it so that it fits within the ethical constraints determined by C , i.e., it is transformed to $\rho_{permissible-ij}$.
2. **Ethical Behavioral Control:** which constrains all active behaviors $(\beta_1, \beta_2, \dots, \beta_m)$ in \mathbf{B} to yield \mathbf{R} with each vector component $\mathbf{r}_i \in P_{permissible}$ set as determined by C , i.e., only lethal ethical behavior is produced by each individual active behavior involving lethality in the first place.
3. **Ethical Adaptor:** if a resulting executed behavior is determined to have been unethical, i.e., $\rho_{ij} \in P_{unethical}$, then use some means to adapt the system to either prevent or reduce the likelihood of

such a reoccurrence and propagate it across all similar autonomous systems (group learning), e.g., an after-action reflective review or an artificial affective function (e.g., guilt).

These architectural design opportunities lie within both the reactive (ethical behavioral control approach) or deliberative (ethical governor approach) components of the hybrid autonomous system architecture. Should the system verge beyond appropriate behavior, after-action review and reflective analysis can be useful during both training and in-the-field operations, resulting in only 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 in a hybrid architecture, similar to the one currently being developed in our laboratory referred to as TAME (for Traits, Attitudes, Moods, and Emotions) [12]. All three of these architectural designs are not mutually exclusive, and indeed can serve complementary roles.

In addition, a crucial design criterion and associated design component, a **Responsibility Advisor**, should make clear and explicit as best as possible, just where *responsibility* vests among the humans involved, should an unethical action be undertaken by the autonomous robot. To do so requires not only suitable training of operators and officers as well as appropriate architectural design, but also an on-line system that generates awareness to soldiers and commanders alike about the consequences of the deployment of a lethal autonomous system. It must be capable to some degree of providing suitable explanations for its actions regarding lethality (including refusals to act). [3] presents the architectural specifications for developing all of the design components above, as shown in Fig. 3.

4. Summary

This paper provides the permeating formalisms for a hybrid deliberative/reactive architecture designed to govern the application of lethal force by an autonomous system to ensure that it conforms with International Law. The details of the proposed architectural design as well as specific recommendations for test scenarios appear in [3]. These efforts are only the first steps in considering an architecture that ensures the ethical application of lethality. It is envisioned that these initial baby steps will lead in the long-term to the development of a system that is

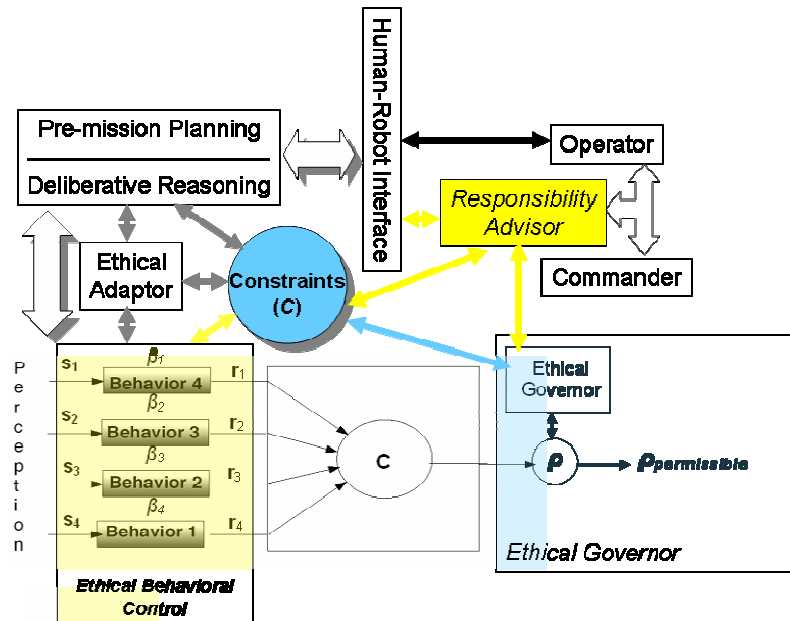


Figure 3: Major Components of an Ethical Autonomous Robot Architecture. The newly developed ethical components are shown in color.

potentially capable of being more humane in the battlefield than humans currently are, and this goal serves as our benchmark for system performance.

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

References

- [1] 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.
- [2] Walzer, M., *Just and Unjust Wars*, 4th Ed., Basic Books, 1977.
- [3] Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, GVI Technical Report, GIT-GVI-07-11, Georgia Tech, 2007.
- [4] Arkin, R.C., *Behavior-based Robotics*, MIT Press, 1998.
- [5] MacKenzie, D., Arkin, R.C., and Cameron, J., 1997. "Multiagent Mission Specification and Execution", *Autonomous Robots*, Vol. 4, No. 1, Jan. 1997, pp. 29-57.
- [6] Balch, T. and Arkin, R.C., "Behavior-based Formation Control for Multi-robot Teams", *IEEE Transactions on Robotics and Automation*, Vol. 14, No. 6, December 1998, pp. 926-939.
- [7] Arkin, R.C., Collins, T.R., and Endo, T., 1999. "Tactical Mobile Robot Mission Specification and Execution", *Mobile Robots XIV*, Boston, MA, Sept. 1999, pp. 150-163.
- [8] Collins, T.R., Arkin, R.C., Cramer, M.J., and Endo, Y., "Field Results for Tactical Mobile Robot Missions", *Unmanned Systems 2000*, Orlando, FL, July 2000.
- [9] Wagner, A., and Arkin, R.C., "Multi-robot Communication-Sensitive Reconnaissance", *Proc. 2004 IEEE International Conference on Robotics and Automation*, New Orleans, 2004.
- [10] Gibson, J.J., *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, MA, 1979.
- [11] Horty, J., *Agency and Deontic Logic*, Oxford University Press, 2000.
- [12] Moshkina, L. and Arkin, R.C., "On TAMEing Robots", *Proc. 2003 IEEE International Conference on Systems, Man and Cybernetics*, Washington, D.C., October 2003.