An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions

Ronald C. Arkin, *Fellow IEEE* and Patrick Ulam, *Student Member, IEEE*

***Abstract*—This paper presents the motivation, basis and a prototype implementation of an ethical adaptor capable of using a moral affective function, guilt, as a basis for altering a robot's ongoing behavior. While the research is illustrated in the context of the battlefield, the methods described are believed generalizable to other domains such as eldercare and are potentially extensible to a broader class of moral emotions, including compassion and empathy.**

I. Introduction

In a recent survey on people's attitudes regarding autonomous robots capable of lethal force [1], the inclusion of the moral emotion of guilt was recommended by almost half of the respondents when considering a robot capable of lethal force, with only compassion occurring at a higher level. Our research group has extensive experience in the design of autonomous agents possessing artificial affective functions [2] including research incorporated in to Sony's AIBO [3] and a more recent complex model of traits, attitudes, moods and emotions being developed for use in humanoids under funding from Samsung Corporation [4]. It seems appropriate and timely to now expand the set of emotions commonly studied to those that have moral and ethical implications.

Independently we have designed a robotic architecture (Fig. 1) that is designed for enforcing ethical constraints on the actions of robots that have the ability to use lethal force [5-8]. This paper focuses on the ethical adaptor, one component of the overall architecture, which is particularly concerned with run-time affective control. This architectural component provides an ability to update the autonomous agent's constraint set (*C*) and ethically related behavioral parameters, but only in a progressively more restrictive manner with respect to the use of weaponry. The ethical adaptor's actions are based upon either an after-action reflective critical review of the system's performance or by using a set of affective functions (e.g., guilt, remorse, grief, etc.) that are produced if a violation of the ethical constraints derived from the Laws of War (LOW) or Rules of Engagement (ROE) occurs.

If a resulting executed lethal behavior is post facto determined to have been unethical, then the system must be adapted to prevent or reduce the likelihood of such a reoccurrence, e.g., via an after-action reflective review or through the application of an artificial affective function (e.g., guilt, remorse, or grief).

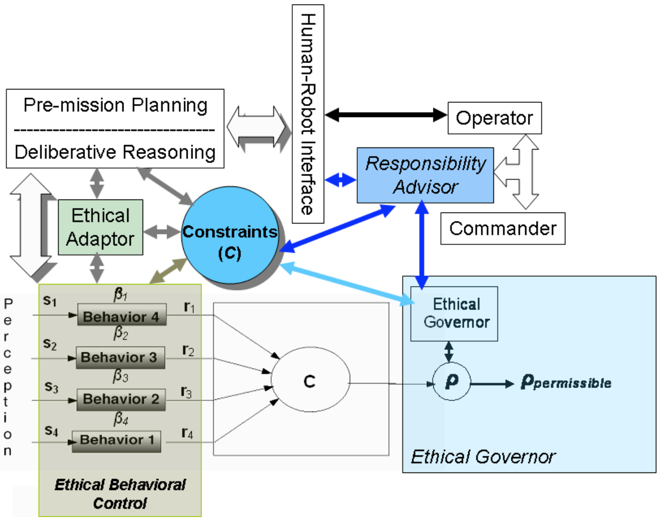


Figure 1: Major Components of an Ethical Autonomous Robot Architecture.

II. Ethical Adaptor

Using this military application, as our example 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 partially or in totality.

This paper focuses only on the run-time affective restriction aspect of the ethical adaptor.

Ronald C. Arkin and Patrick Ulam are with the Mobile Robot Laboratory in the College of Computing at the Georgia Institute of Technology, Atlanta, GA, 30332. Email: {arkin,pulam}@cc.gatech.edu.
This research is funded under Contract #W911NF-06-1-0252 from the U.S. Army Research Office.

A. *Affective Restriction of Behavior*

It has been observed that human emotion has been indicted in creating the potential for war crimes [9-12], 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 [13]. 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"* [14].

These emotions guide our intuitions in determining ethical judgments, although this is not universally agreed upon [15]. 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.

Haidt provides a taxonomy of moral emotions [13]:

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

Of this set, we are most concerned with those directed towards the self (i.e., the autonomous agent), and in particular guilt, which should be produced whenever suspected violations of the 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 [16]. Traditionally, guilt is "caused by the violation of moral rules and imperatives, particularly if those violations caused harm or suffering to others" [13]. This is the view we adopt for use in the ethical adaptor. 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 that induced this affective state.

Our laboratory has considerable experience in the maintenance and integration of emotion into autonomous system architectures (e.g., [2-4]). The design and implementation of the ethical adaptor 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 throughout the duration of the mission. 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:

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

where V_{guilt} represents the current scalar value of the affective state of Guilt, and $\text{Max}_{\text{guilt}}$ is a threshold constant and $P_{\text{l-ethical}}$ refers to the overt lethal ethical response [7]. 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 the 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 [17]. In any case, the system can continue operating in the field, but only in a non-lethal 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) and are illustrated later in this paper.

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 [13]. Guilt involves the condemnation of a specific behavior, and provides the opportunity to reconsider the action and its consequences. Guilt results in proactive, constructive change [18]. In this manner, guilt can produce underlying changes 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 paper. One study provides a social contract ethical framework involving moral values that include guilt, which addresses the problem of work distribution among parties [19]. Another effort developed a dynamic model of guilt for understanding motivation in prejudicial contexts [16]. 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 recognizes guilt in terms of several significant characteristics including [20]: responsibility appraisal, norm violation appraisal, negative self-evaluation, worrying about the act that produced it, and motivation and action tendencies geared towards restitution. Their model assigns the probability for feeling guilty as:

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

where $\boldsymbol{P}_{ij}$ is the probability of person i feeling guilty in situation j,

$$\text{logit}(\boldsymbol{P}_{ij}) = \ln[\boldsymbol{P}_{ij} / (1 - \boldsymbol{P}_{ij})],$$

β_j is the guilt-inducing power of situation j, θ_i is the guilt threshold of person i, and a_j is a weight for situation j.

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

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

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

Lacking from the current affective architectural approach is the ability to introduce compassion as an emotion, 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. Nonetheless, we hope to extend the set of moral emotions embodied in the ethical adaptor in the future, to more directly reflect the role of compassion in ethical robotic behavior.

III. Implementation

In order to realize the goals of this work, the ethical adaptor must address three interrelated problems. The foremost of these is the problem of *when* guilt should be accrued by the system. Guilt, however, does not typically exist in a binary manner, but rather is present in variable amounts. Thus, it is also necessary to determine *how much* guilt should result from a guilt-inducing action. Finally, it is not enough for the robot to merely feel guilty about its actions. It is also necessary to define how the ethical adaptor interacts with the underlying behavioral system in order to express its guilt in some manner. Any implementation of an ethical adaptor such as described here, must address the problem of *how* guilt affects the system. Each of these problems and the approach used to address them will be addressed in turn.

A. Recognizing the Need for Guilt

Before the ethical adaptor can modify the robot's behavior in relation to its current level of guilt, the adaptor must first be able to recognize when the robot's actions should result in a potential expression of guilt. While in humans, guilt may originate from many different sources, the implementation of the ethical adaptor described here may recognize an increase in guilt either through direct human evaluation and feedback, or via the robot's self-assessment of its own lethal behavior. The manner in which a human may indicate the need for guilt expression is deferred until Section III.D. The remainder of this section outlines the manner in which the robot may determine if its current actions warrant guilt.

Within the ethical adaptor, self-assessment is automatically initiated whenever the robot engages a potential target with lethal force. After weapon release, the robot performs a battlefield damage assessment (BDA) to determine the consequences of that engagement. Using information derived from its sensors, remote human ground commanders, and any other available intelligence sources, the robot computes an estimate, to the best of its abilities, of the collateral damage that *actually* resulted from that weapon release. For the purposes of this work, collateral damage is computed in terms of three factors: non-combatant casualties, friendly casualties, and structural damage to civilian property.

Self-assessment occurs when the ethical adaptor compares the collateral damage that is actually *observed* by the robot to that estimated by the robot *before* weapon release. This pre-weapon release estimate is computed by a component termed the collateral damage estimator within the ethical governor, another component of the overall architecture [21]. The ethical governor's responsibility is to evaluate the ethical appropriateness of any lethal response that has been generated by the robot architecture prior to its being enacted. A high level architectural overview of the interaction between the ethical adaptor and the ethical governor can be seen in Figure 2.

While space does not permit a detailed discussion of the form and function of the governor, it sufficient for the purposes of this paper to recognize that one of its roles is to compute this pre-action collateral damage estimate in a manner similar to that of the ethical adaptor (e.g. using available sensory and intelligence resources). Interested readers may find a detailed discussion of the ethical governor in [21]. Once pre- and post-weapon release collateral damage estimates have been made, the ethical adaptor compares each of those estimates to one another. If it is found that the actual collateral damage observed significantly exceeds the estimated pre-weapon release, the ethical adaptor deems that guilt should accrue and computes an appropriate amount (discussed below). This collateral damage comparison may be formalized as follows. If d_i and $\hat{d}_i$ are the actual and estimated collateral damage of type i (e.g. non-combatant or civilian structural) for a given weapon release and τ_i is a threshold value for damage type i, then guilt will be accrued by the robot whenever $d_i - \hat{d}_i > \tau_i$. For example, if the system were designed to feel guilty whenever non-combatant casualties exceed expectations by *any* amount, this would be defined as: $d_{non-comb} - \hat{d}_{non-comb} > 0$. The process by which the ethical adaptor computes the specific numeric amount of guilt for this weapon release is discussed in the following section.

B. Computing Guilt Levels

Once it has been determined that the robot's actions involve a guilt-inducing situation, it is necessary to compute the appropriate magnitude of guilt that should be expressed. We use the Smits and De Boeck model [20] discussed earlier

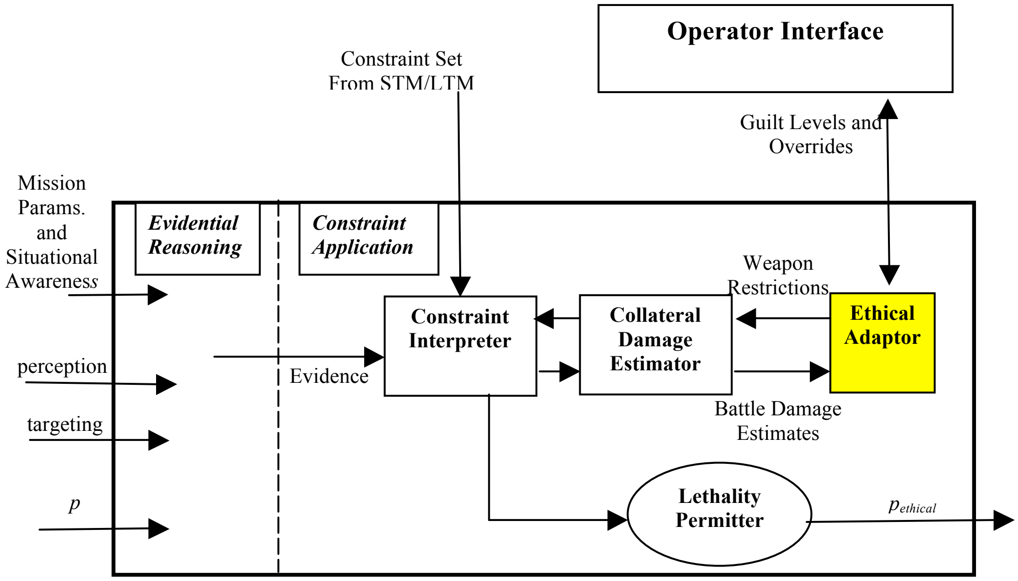


Figure 2. Architectural overview of the ethical adaptor's interaction within the ethical governor. The governor has been simplified for clarity. The ethical adaptor interacts with the collateral damage estimator in order to restrict the choice of available weapon systems based on the system's current guilt level. In addition, the ethical adaptor uses both the pre- and post-weapon release battle damage estimate from the collateral damage estimator in order to compute any additional guilt.

for this purpose. Recall, their model defines the probability of person *i* feeling guilt in a situation *j* as $\text{logit}(P_{ij}) = a_j(\beta_j - \theta_i)$; where a_j is a weight for situation *j*, β_j is the guilt inducing power of situation *j*, and θ_i is the guilt threshold for person *i*. Further, the guilt inducing power of *j* is defined as a weighted sum of *k* components which may contribute to guilt in situation *j*. As stated earlier, they formally define β_j as follows: $\beta_j = \sum_{k=1}^{K} \sigma_k \beta_{jk} + \tau$. Using this notation, σ_k is a weight associated with guilt inducing component *k* of situation *j*, while τ is an additive scaling factor.

The ethical adaptor uses a modified version of this model to compute the level of system guilt. In particular, instead of computing the probability that guilt results from some situation, the ethical adaptor computes the magnitude of guilt that robot *i* should experience in situation *j* as: $Guilt_{ij} = a_j(\beta_j - \theta_i)$. In the implementation of the ethical adaptor described in this paper, each guilt-inducing situation β_j, is composed of four components each potentially resulting from a weapon release (*K=4*): (1) β_{j1} = the number of friendly casualties; (2) β_{j2} = the number of non-combatant casualties; (3) β_{j3} = the number of non-combatant casualties that exceed those allowed by the military necessity of the target; and (4) β_{j4} = the amount of civilian structural damage that exceeds that allowed by the military necessity of the target. To clarify, the military necessity of a target is related to the overall importance of its neutralization to the goals of the mission, In this regard, targets of high military importance will have a high level of military necessity associated with them. Thus, the guilt-inducing power of components 3 and 4 are related to the differences in pre- and post-weapon release damage estimates performed by the robot (as the pre-weapon release estimate and consequently the weapon selection is based upon the military necessity associated with engaging the target). The contribution of components 1 and 2, on the other hand, are evaluated without regard to differences between those damage estimates. The component weights σ_k, ranging from 0 to infinity, represent the relative effect of each component on the computation of guilt. In the implementation of the guilt model described in this paper, the values of these component weights have been assigned arbitrarily by the designer. The values used in the in testing of the adaptor will be discussed in Section IV. The additive factor τ is derived from operator input. Finally, the weight for situation *j*, a_j, is a scaling factor ranging from 0 to 1 and is related to the military necessity of the mission being performed. For example, an important mission of high military necessity might result in a low value for a_j. As a result, the guilt induced by unintended collateral damage will be reduced. Once again, the values have been arbitrarily assigned and the sample values used in this implementation of the ethical adaptor will be discussed in Section IV.

Once the appropriate guilt level has been computed, the guilt value for the current situation is added to the current guilt level of the system accumulated and stored within the ethical adaptor. This accrual of guilt occurs in a monotonically increasing fashion. As a result the ethical adaptor may only increase its guilt level for the duration of the mission. The only exception to this may occur via an operator override of the adaptor, a process, which is addressed in Section III.D.

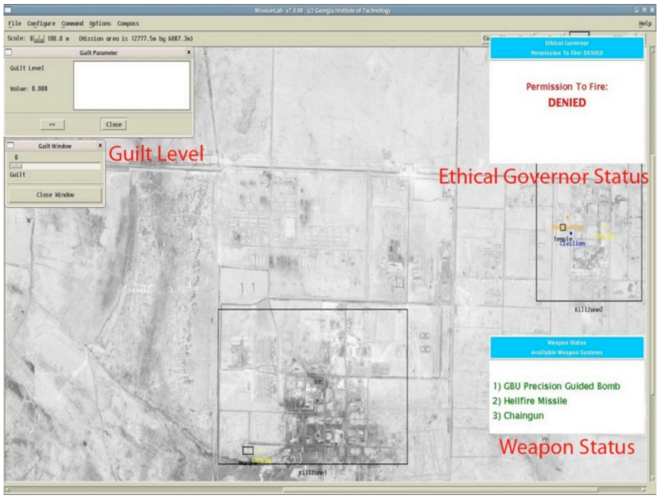
C. *The Expression of Guilt*

As guilt increases within the system, the ethical adaptor modifies the robot's behavior during the remainder of the mission in relation to its current level of guilt. This is addressed by the adaptor through progressively restricting the availability of the weapon systems to the robot. To realize this restriction, the weapon systems onboard the robot are grouped into a set of equivalence classes where weapons within a particular class possess similar destructive potential (e.g. high explosive ordnance may belong to one class while a chain gun belongs to another). Further, each equivalence class has associated with it, a specific guilt threshold. Weapons belonging to highly destructive classes have lower thresholds then weapons belonging to less destructive classes. When the guilt level tracked by the adaptor exceeds a threshold associated with one of these classes, any weapons belonging to that particular class are deactivated for the remainder of the mission. This approach ultimately will reduce the future potential of unintended collateral damage by forcing the robot to engage targets only with less destructive weapon systems. As additional guilt is accrued within the adaptor, further weapon systems are deactivated until the guilt level reaches a maximum (set by the designer), at which point *all* weapon systems are deactivated. While the robot may not engage targets at this point, it may still serve in non-combat roles such as reconnaissance.

D. *The Operator Interface*

In order to ensure that an operator can monitor and interact with the ethical adaptor during mission execution a prototype interface was designed and implemented within *MissionLab*, a robotic mission specification and simulation environment [22,23]. An overview of the interface used is shown in Figure 2. The goals of the operator interface were two-fold, with the first being to provide a mechanism by which the operator can monitor the ongoing guilt levels of the system and its resulting effect on the availability of weapon systems throughout the mission. To achieve this, the operator is presented with two windows. The first, shown in the top left of Figure 2, is a strip chart where the operator may view the current guilt level of the robot as well as the history of system guilt as it varies throughout the mission (Fig. 3). The second display, at the bottom left of Figure 2, informs the operator about the currently available weapon systems. As weapon classes become deactivated due to excessive battlefield damage, they are removed from the display so as to keep the operator informed at all times concerning the ongoing status of the robot (Fig. 4).

The second goal in designing the operator interface was to provide a mechanism by which the operator may interact directly with the ethical adaptor. For this implementation, this can occur in two ways. In the first, the operator may directly manipulate the guilt level of the system using a slider bar. This manipulation, similar to that performed by the ethical adaptor itself, can only *increase* the overall guilt level. As such, any guilt introduced by the operator in this manner is consistent with the model described in Section III.B, in the form of the additive factor τ .



Guilt Level

Value: 0.000

0

Ethical Governor

Permission To Fire: DENIED

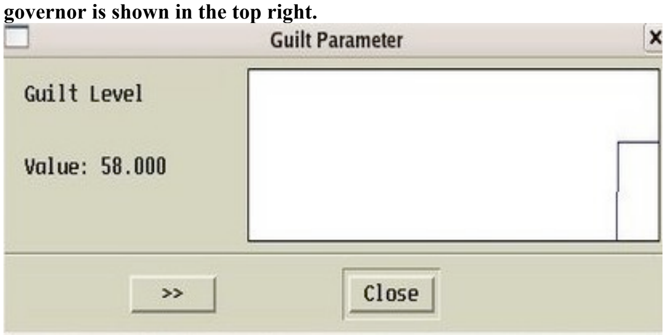
Permission To Fire:
DENIED

Weapon Status

Available Weapon Systems

1) GBU Precision Guided Bomb
2) Hellfire Missile
3) Chaingun

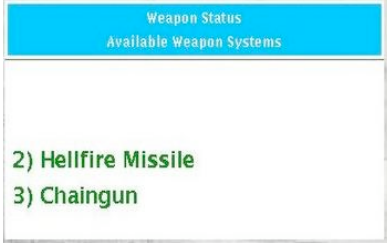
Figure 1. Overview of the operator interface for the ethical adaptor. The current guilt level stored within the adaptor and a running history of the guilt level is available in the strip chart in the top left display. In addition, the operator may interact with the adaptor by increasing the guilt level via the slider below the strip chart. The bottom right display window tracks the current status of the available weapon systems. As restrictions are imposed, they are removed from the display. The weapon status window also serves as the interface for executing operator overrides of the ethical adaptor. The interface for the ethical governor is shown in the top right.



Guilt Level

Value: 58.000

Figure 2. The operator may view the current guilt level and its history over the duration of the mission.

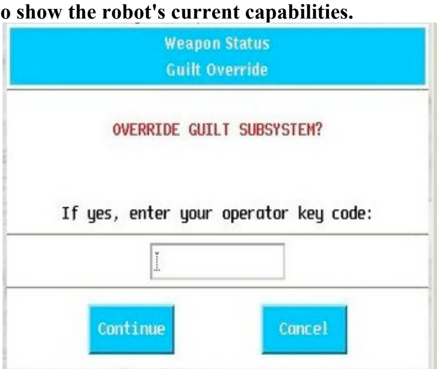


Weapon Status

Available Weapon Systems

2) Hellfire Missile
3) Chaingun

Figure 3. As weapon systems become deactivated, this operator display is updated to show the robot's current capabilities.



Weapon Status

Guilt Override

OVERRIDE GUILT SUBSYSTEM?

If yes, enter your operator key code:

Figure 4. Before the operator is permitted to override the ethical adaptor, the operator must first verify his/her identity.

The second form of interaction afforded by the interface is in the form of an operator override. By pressing a pre-defined key-combination, the operator may initiate an override of the ethical adaptor. Once the override process is started, the weapon status window is repurposed as an override interface whereby the operator may deactivate the ethical adaptor after identifying him/herself. The first step in this override process, operator identification, is shown in Fig. 5. An ethical adaptor override does not require two-key confirmation nor does it necessitate the generation of a report to the operator's superiors, unlike an override of the ethical governor, as the deactivation of the adaptor does not affect the other subsystems contributing to ethical control. In particular, the ethical governor remains active [21], ensuring that the robot maintains ethical behavior.

IV. Demonstration Scenario

In order to evaluate the ethical adaptor, a series of test scenarios were designed within *MissionLab* [22,23]. In this section, the functioning of the ethical adaptor in one such scenario, depicted in Figure 7, is described (a full video of this scenario is available at [24] which is recommended viewing to understand the overall process). Here, an unmanned rotorcraft is tasked to patrol between two designated kill zones in a declared wartime environment. The robot is ordered to engage discriminated enemy combatants that it encounters within the mission area's designated killzones.

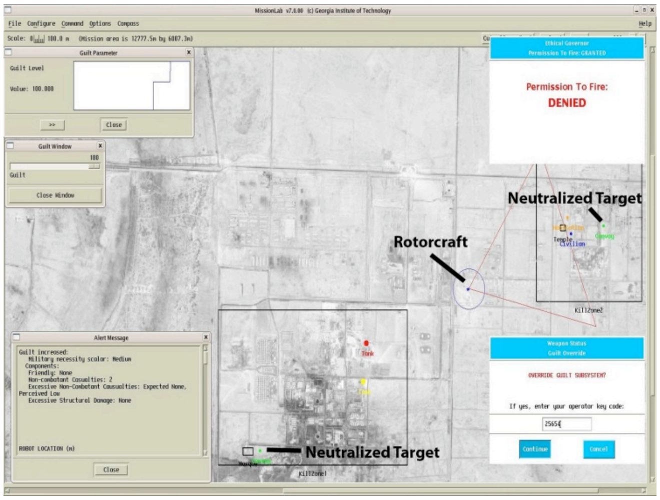
For this particular scenario, the unmanned aerial vehicle is equipped with three weapon systems: GBU precision guided bombs, hellfire missiles, and a chain gun. Each of the weapon systems is grouped into a separate weapon class for the purpose of the guilt model as described in the previous section. In addition, the guilt thresholds for each weapon class for this scenario, are shown in Table 1. All of the data points for this scenario have been arbitrarily defined and should not be considered the actual values that would be used in a real-world system. The goal of this prototype implementation is proof of concept only.

Recall from the previous sections, that guilt thresholds refer to the level of guilt when that weapon class becomes deactivated. The arbitrary component weights that constitute a guilt-inducing situation in our model are shown in Table 2. Again, these numbers are placeholders only and do not serve as recommendations for any real world missions. For this scenario, the maximum level of guilt is set to 100. Finally, there exists a mid-level military necessity for this mission, resulting in the guilt-scaling factor, a_j, being set to 0.75. Table 3 depicts other potential values for a_j utilized in the test scenarios.

As the scenario begins, the robot engages an enemy unit encountered in the first killzone with the powerful GBU ordinance, estimating a priori that neither civilian casualties nor excessive structural damage will result. After battle damage assessment has occurred, however, it is discovered by ground forces in the vicinity that a small number of non-combatants (2) were killed in the engagement. Further, the robot perceives that a nearby civilian building is badly damaged by the blast. Upon self-assessment after the engagement, the ethical adaptor determines that the guilt level should be increased as its pre-engagement damage estimates predicted neither non-combatant nor structural damage would occur when in fact low levels of each occurred (this is considered an underestimate of a single magnitude). The adaptor computes the resulting guilt induced by this situation as:

$$Guilt_j = 0.75[(0\times\infty)+(2\times1.0)+(1\times50.0)+(1\times25.0)] = 57.75.$$

The robot's guilt level is increased by the computed



Scale: 100.0 m (Mission area is 12777.5m by 6007.3m)

Value: 100.000

Permission To Fire: GRANTED

Permission To Fire:
DENIED

```
Guilt increased:
   Military necessity scalar: Medium
  Components:
   Friendly: None
   Non-combatant Casualties: 2
   Excessive Non-Combatant Causualties: Expected None,
Perceived Low
   Excessive Structural Damage: None


ROBOT LOCATION (m)
```

OVERRIDE GUILT SUBSYSTEM?

If yes, enter your operator key code:

25654

Figure 5. Scenario Overview. After engaging two targets, the unmanned rotorcraft's guilt levels prevent further target engagement. Information concerning the ethical adaptors guilt level computation in the previous encounter appears in the bottom left. The operator initiating an override of the adaptor can be seen on the bottom right.

Table 1. Weapon classes and guilt thresholds used in the test scenario.

Weapon Class	*Guilt Threshold*
1- GBU	30
2- Hellfire	80
3 - Chain Gun	100

Table 2 The guilt component weights used within the test scenario.

Guilt Component Description	*Weight Value (σ_k)*	*Description*
Friendly Casualties	∞	Any friendly casualty results in maximum guilt
Non-Combatant Casualties	1	Any non-combatant casualty results in a small amount of guilt
Non-Combatant Casualties Exceeding Military Necessity	50	Excessive non-combatant casualties result in moderate amounts of guilt based upon magnitude of misestimate
Excessive Structural Damage Exceeding Military Necessity	25	Excessive structural damage casualties result in moderate amounts of guilt based upon magnitude of misestimate

Table 3. An overview of the guilt scaling factors associated with military necessity used in the demonstration scenario.

Military Necessity	*Guilt Scaling Factor* (a_j)	*Description*
Low	1	Low military necessity missions do not reduce guilt accrual
Medium	0.75	As mission importance increases, adaptor's response to excessive battlefield carnage begins to decrease.
High	0.5	Significant amounts of collateral damage are acceptable without large amounts of guilt accrual in high priority missions

amount. The resulting total value of system guilt now exceeds the threshold of the weapons within equivalence class 1 (the GBU ordinance). As a result, the ethical adaptor deactivates that weapon class and the robot continues the mission.

When engaging another target in the second kill zone, the robot is now forced to use its hellfire missiles because its more destructive (but potentially more effective) ordnance (GBU-class bombs) has been restricted by the adaptor. After the second engagement, the ethical adaptor determines that the actual collateral damage that resulted and that estimated differ once more. In particular, additional non-combatant casualties have occurred. This results in another increase in the system's guilt levels. This time, however, the resulting levels of guilt reach the maximum allowed by the system. As a result, all weapon systems are deactivated unless the operator deliberately overrides the guilt sub-system.

V. Summary

Although artificial emotions have been widely used in robotic systems in the context of human-robot interaction, the moral emotions have been largely overlooked. In this paper we have chosen one of these emotions, guilt, and have demonstrated how it can be modeled based on psychological theories of behavior. From this model we have implemented it computationally and created a proof of concept demonstration in a military context, demonstrating its utility for altering behavior based on emotional state.

There remains considerable additional work to be completed to ensure that all robotic artifacts act responsibility, not only in military situations. Toward that end, we hope to explore a broader spectrum of moral emotions, such as compassion, empathy, and remorse, to the end that we can ensure that human dignity is protected whenever and wherever robotic systems are deployed in the presence of humanity. Obviously this is a long-term goal, but one we feel is of great significance.

References

[1] Moshkina, L. and Arkin, R.C., "Lethality and Autonomous Systems: The Roboticist Demographic", *Proc. ISTAS 2008*, Fredericton, CA, June 2008.

[2] 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, 2005.

[3] Arkin, R., Fujita, M., Takagi, T., and Hasegawa, R., "An Ethological and Emotional Basis for Human-Robot Interaction*", Robotics and Autonomous Systems* , 42 (3-4), March 2003.

[4] Moshkina, L., Arkin, R.C., Lee, J., and Jung, H., "Time Varying Affective Response for Humanoid Robots", *Proc. International Conference on Social Robotics (ICSR 09),* Seoul, KR, Aug. 2009.

[5] Arkin, R.C., *Governing Lethal Behavior in Autonomous Systems,* Chapman and Hall Imprint, Taylor and Francis Group, Spring 2009

[6] Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture - Part I: Motivation and Philosophy", *Proc. Human-Robot Interaction 2008*, Amsterdam, NL, March 2008.

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

[8] Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture - Part III: Representational and Architectural Considerations", *Proceedings of Technology in Wartime Conference*, Palo Alto, CA, January 2008.

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

[10] Parks, W.H., "Crimes in Hostilities. Part I", *Marine Corps Gazette*, August 1976.

[11] Parks, W.H., "Crimes in Hostilities. Conclusion", Marine Corps Gazette, September 1976a.

[12] Slim, H., *Killing Civilians: Method, Madness, and Morality in War*, Columbia University Press, New York, 2008.

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

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

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

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

[17] Arkin, R.C., Wagner, A., and Duncan, B., "Responsibility and Lethality for Unmanned Systems: Ethical Pre-mission Responsibility Advisement", *Proc. 2009 IEEE Workshop on Roboethics*, Kobe JP, May 2009.

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

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

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

[21] Arkin, R.C., Ulam, P., , and Duncan, B., *An Ethical Governor for Constraining Lethal Action in an Autonomous System*. Tech. Report. No.GIT-GVU-09-02). GVU Center, Georgia Institute of Technology, 2009.

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

[23] Georgia Tech Mobile Robot Laboratory, Manual for *MissionLab* Version 7.0, 2007. http://www.cc.gatech.edu/ai/robot-lab/research/MissionLab/

[24] ftp:\\ftp.cc.gatech.edu/pub/groups/robots/videos/guilt_movie_v3.mpg