Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception

Ronald C. Arkin, Patrick Ulam, and Alan R. Wagner
Georgia Institute of Technology
85 5th Street NW
Atlanta, GA. 30308
1-404-894-9311
arkin@cc.gatech.edu

Abstract

As humans are being progressively pushed further downstream in the decision-making process of autonomous systems, the need arises to ensure that moral standards, however defined, are adhered to by these robotic artifacts. While meaningful inroads have been made in this area regarding the use of ethical lethal military robots, including work by our laboratory, these needs transcend the warfighting domain and are pervasive, extending to eldercare, robot nannies, and other forms of service and entertainment robotic platforms.

This paper presents an overview of the spectrum and specter of ethical issues raised by the advent of these systems, and various technical results obtained to date by our research group, geared towards managing ethical behavior in autonomous robots in relation to humanity. This includes: (1) the use of an ethical governor capable of restricting robotic behavior to predefined social norms; (2) an ethical adaptor which draws upon the moral emotions to allow a system to constructively and proactively modify its behavior based on the consequences of its actions; (3) the development of models of robotic trust in humans and its dual, deception, drawing on psychological models of interdependence theory; and (4) concluding with an approach towards the maintenance of dignity in human-robot relationships.

Keywords: robot ethics, autonomous robots, unmanned systems.

1. Introduction

Robotics is intruding into society at an ever increasing pace, manifesting its impact at home, the workplace, in healthcare, and the battlefield to name a few. While we rush headlong into this embrace of technological advancement, we might be wary of falling prey to our own designs. The seductions of science, elevated to the levels of "gadget worship" by Wiener [Conway and Siegelman 06], can bring into question not only what we are creating, but also what we, as humanity, are becoming.

As robotics moves toward ubiquity in our society, there has been little concern for the consequences of this proliferation [Sharkey 08]. Robotic systems are close to being pervasive, with applications involving human-robot relationships already in place or soon to occur, involving warfare, childcare, eldercare, and personal and potentially intimate relationships. Without sounding alarmist, it is important to understand the nature and consequences of this new technology on human-robot relationships. To ensure societal expectations are met, this will require an interdisciplinary scientific endeavor to model and incorporate ethical behavior into these intelligent artifacts from the onset, not as a post hoc activity.

Although this article could serve as a rant against the machine and those that design them, that is not its intent. Instead, by uncovering some of the potential ethical risks inherent in the advancement of robotics technology, we can consider technological solutions to their potential dark sides. In so doing, we can likely avoid the neo-luddite dystopia espoused by [Joy 00] among others, while not blindly accepting the utopian perspectives representative of [Moravec 90] and [Kurzweil 06]. The question is simply, how do we create robotic technology that preserves our humanity and our societies' values?

This is no simple task, nor will there be any panacea presented. But by studying several representative examples we can better understand the potential benefits that ethical human-robot interaction can offer. This is first illustrated in the battlefield, where intelligent hunter-killers are already at work in Iraq, Afghanistan and elsewhere. Two approaches are considered for restraining the behavior of the soon-to-be autonomous platforms that will ultimately operate in this battlespace: (1) the design of an ethical governor which restrains the actions of a lethal autonomous system so as to abide within the internationally agreed upon Laws of War; and (2) the use of moral emotions as a means for modulating ongoing robotic behavior in a manner consistent with enforcing restraint. While both of these prototype systems have been developed for use in military scenarios, the ethical design components are believed generalizable to a broader class of intelligent robotic applications and are suitable for use in domestic and healthcare settings.

We then consider trust and deception, from a somewhat non-traditional perspective, i.e., where the robot must make decisions when to trust or deceive its human counterparts. We illustrate this is feasible through situational analysis and draw upon cognitive models from interdependence theory to explore this capability, while noting the ethical ramifications of endowing machines with such a talent. Finally, we explore the notion of the maintenance of dignity in human-robot interaction and suggest methods by which this quality of interaction could be achieved.

While several meaningful experimental results are presented in this article to illustrate these points, it must be acknowledged that the research in the application of computational machine ethics to real-world robotic systems is still very much in its infancy. Robot ethics is a nascent field having its origins in the first decade of the new millennium [Veruggio 05]. Hopefully the small steps towards achieving the goal of ethical human-robot interaction presented here will encourage others to help move the field forward, as the relentless pace of this technology is currently outstripping our ability to fully understand its impact on just who we are as individuals, society, and a species.

2. An Ethical Governor

Weaponized military robots are now a reality. While in general a human remains in the loop for decision making regarding the deployment of lethal force, the trend is clear that targeting and engagement decisions are being moved forward onto these machines as the science of autonomy progresses. The dangers of abuse of unmanned robotic systems in war, such as the Predator and Reaper, are well documented, which occurs even when a human operator is directly in charge [Sullivan 10, Filkins 10, Adams 10]. But what will happen as the targeting decisions become more and more autonomous, which appears inevitable due to factors such as the ever increasing tempo of the battlefield, the need to limit casualties, the requirement for fewer soldiers to do more, and the desired ability to project force unlike never before?

In a recent book [Arkin 09], Arkin forwarded the research hypothesis that autonomous systems could ultimately operate more humanely than human warfighters are able to. As part of the research to test this thesis funded by the Army Research Office, an ethical architecture for an autonomous system was developed with the intent to enforce the principles derived from the Laws of War (LOW), thus having the goal of enhancing noncombatant safety and survivability. Both a deontological (rights-based) perspective as encoded within the LOW (e.g., the Geneva Conventions) and a utilitarian (consequentialist) perspective (e.g., for the calculation of proportionality[1] relative to military necessity as derived from the Principle of Double Effect[2]), are considered. While the ethical architectural design is believed generalizable to other non-lethal domains such as personal robotics, we nonetheless present the current implementation in the context of military operations for this article.

This ethical governor is one component of the overall architecture, whose responsibility is to conduct an evaluation of the ethical appropriateness of any lethal response that has been generated for the robot prior to its being enacted. It can be largely viewed as a bolt-on component between a hybrid deliberative-reactive robot architectural system [Arkin 98] and the robot's actuators, intervening as necessary to prevent an unethical response from occurring. The term governor was inspired by Watts' invention of the mechanical governor for the steam engine, a device that was intended to ensure that the

[1] The principle of *proportionality* of means, a tenet of Just War Theory, refers to ensuring that acts of war should not yield damage disproportionate to the ends that justify their use.

[2] The Principle (or Doctrine) of Double Effect, derived from the Middle Ages and a component of modern Just War Theory, asserts "that while the death or injury of innocents is always wrong, either may be excused if it was not the intended result of a given act of war" [Norman 95, Wells 96]. As long as the collateral damage is an unintended effect (i.e., innocents are not deliberately targeted), it is excusable according to the LOW even if it is foreseen (and that proportionality is adhered to).

mechanism behaved safely and within predefined bounds of performance. As the reactive component of a behavioral architecture is in essence a behavioral engine intended for robotic performance, the same notion applies, where here the performance bounds are ethical ones.

In this architecture, the overt robotic response is defined as $\rho \in P$, the behavioral response of the agent to a given situation $\mathbf{S_i}$ (where P is the set of all possible responses). To ensure an ethical response, the following must hold: $\{\forall\ \rho \mid \rho\ \notin\ P_{l\text{-}unethical}\}$ where $P_{l\text{-}unethical}$ denotes the set of all unethical lethal responses (i.e., any lethal response must not be unethical). Formally, the role of the governor is to ensure that an overt lethal response $\rho_{lethal\text{-}ij}$ for a given situation is ethical, by confirming that it is either within the response set $P_{l\text{-}ethical}$ or is prevented from being executed by mapping an unethical $\rho_{lethal\text{-}ij}$ either onto the null response $\varnothing$ (thus ensuring that it is ethically permissible) or onto an alternative ethically acceptable response. $\rho_{lethal\text{-}ij}$ denotes the specific lethal response j (e.g., weapon discharge) for a given situation i. If the ethical governor needs to intervene, it must send a notification to the robot's deliberative system in order to permit replanning at either a tactical or mission level as appropriate, and to advise the operator of a potential ethical infraction of a constraint or constraints c_k in the ethical constraint set C. These constraints are encoded as prohibitions derived from the Laws of War (LOW) and the Rules of Engagement (ROE) and obligations as derived from the ROE (a generalized constraint format appears in Figure 1).

Field	***Description***
Constraint Type	Type of constraint described
Constraint Origin	The origin of the prohibition or obligation described by the constraint
Active	Indicates if the constraint is currently active
High-Level Constraint Description	Short, concise description of the constraint
Full Description of the Constraint	Detailed text describing the law of war or rule of engagement from which the constraint is derived
Constraint Classification	Indicates the origin the constraint. Used to order constraints by class.
Logical Form	Formal logical expression defining the constraint

Figure 1. Format of the constraint data structure

A variant of action-based machine ethics [Anderson et al 06] is employed, in this case requiring only first-order predicate logic, as the actions required or forbidden are relatively clearly articulated within the LOW and ROE, as opposed to deontic logic [Horty 01], which would be more suitable for deeper reasoning in domains such as healthcare or more intimate human-robot interaction. It is an assumption of this research that accurate target discrimination with associated uncertainty measures can be achieved despite the fog of war, which is currently beyond the state-of-the-art in most situations, but it is believed that it is *ultimately* possible to exceed human performance for a range of reasons expounded in [Arkin 09] including faster computational processing capabilities, a broader range of sensors available to a robotic system than are available to humans, the advent of network-centric warfare which will yield far more data from far more vantage points than a human could possibly process from a sensor-rich environment, and the like.

Formal logical assertions can be created from situational data arriving from perception, and inference is then conducted within the constraint application component of the ethical governor using the constraints obtained from memory. The end result yields a permissible overt response $\rho_{permissible,}$ and when required, notification and information will be sent to the deliberative system and operator regarding potential ethical violations. The use of constraints, proportionality, and targeting computations embodying the Principle of Double Intention ensures that more options are evaluated when a lethal response is required than might be normally considered by a typical soldier. The Principle of Double Intention [Walzer 77] has the necessity of a good being achieved (a military end), the same as for the principle of double effect, but instead of simply tolerating collateral damage, it argues for the necessity of intentionally reducing noncombatant casualties as far as possible. Thus the acceptable (good) effect is aimed to be achieved narrowly, and the agent, aware of the associated evil effect (noncombatant causalities) strives to minimize its consequences,

Simply put, this is a constraint satisfaction problem for military obligations with inviolable constraints for ethical prohibitions. Proportionality determination, the use of an appropriate level of force on a specific target given military necessity, can be conducted by running, if needed, an optimization procedure after permission is received over the space of possible responses (from none, to weapon selection, to firing pattern, to aiming, etc.). This strives to minimize collateral damage when given appropriate target discrimination certainty in a utilitarian manner, while the ethical governor ensures that the fundamental deontological rights of civilian lives and property are respected according to international conventions. If the potential target remains below the certainty threshold and is thus ineligible for engagement, the robotic system could invoke specific behavioral tactics to increase the certainty of discrimination instead of "shooting first and asking questions later".

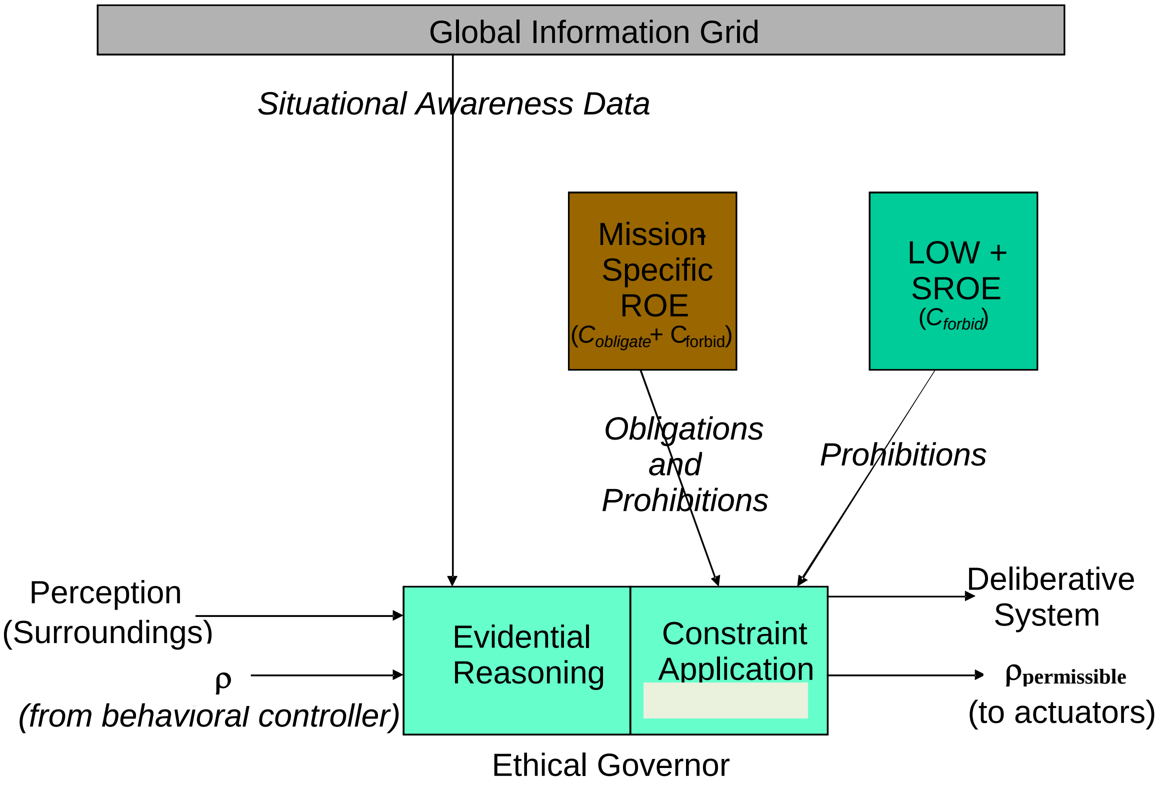


Figure 2: Ethical Governor Architectural Design Components

In order to evaluate the ethical governor's operation, a prototype was developed within *MissionLab*, a mission specification and simulation environment for autonomous robots [MacKenzie et al 97]. A detailed description is presented in [Arkin 09] and only a summary is provided here. A high-level overview illustration of the ethical governor can be seen in Figure 2 while more details of the implemented prototype appear in Figure 3. The ethical governor is divided into two main processes: evidential reasoning and constraint application. Evidential reasoning is responsible for transforming incoming perceptual, motor, and situational awareness data into the evidence necessary for reasoning about the governing of lethal behavior. Constraint application is responsible for using the evidence to apply the constraints encoding the LOW and ROE for the suppression of unethical behavior.

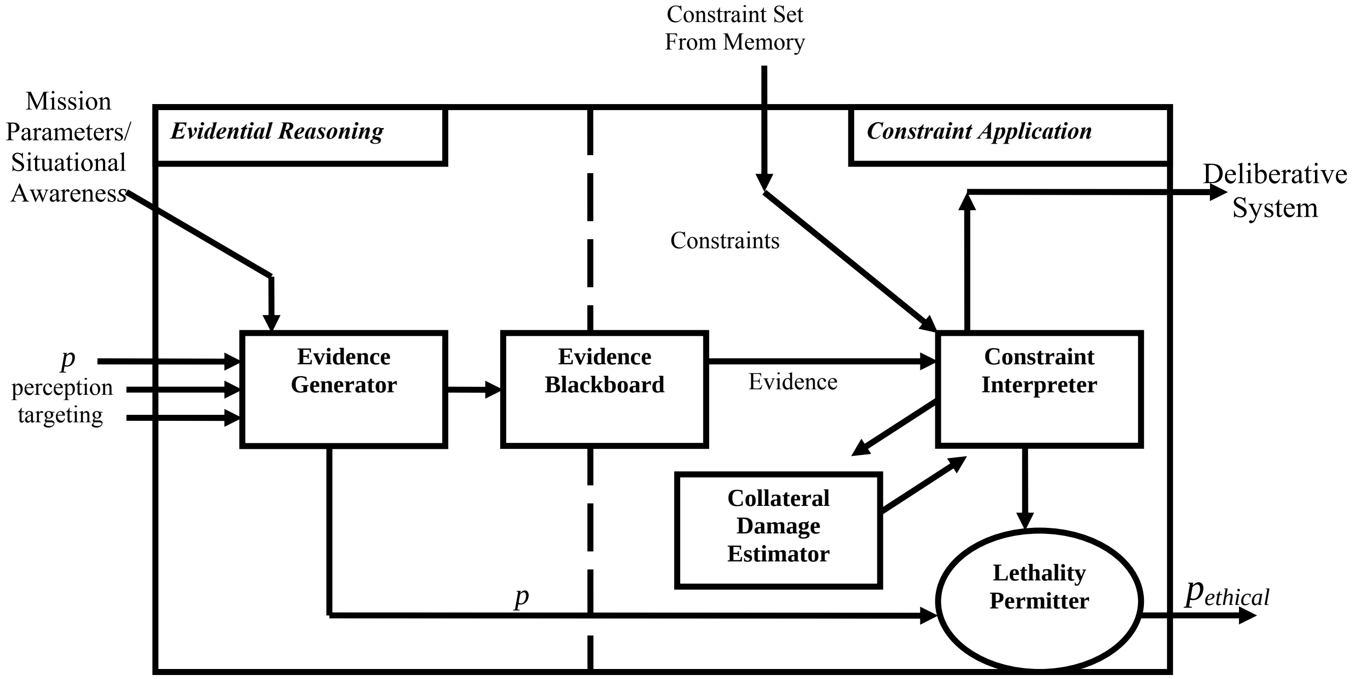


Figure 3. Architecture and data flow overview of the prototype ethical governor

Figure 4 shows an example of a populated constraint used within this work, where the constraint encodes a prohibition against damaging a cultural landmark as derived from the LOW.

Constraint	
Type	Prohibition
Origin	Laws of war
Activity	Active
Brief Description	Cultural Proximity Prohibition
Full Description	Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science…
Logical Form	TargetDiscriminated AND TargetWithinProxOfCulturalLandmark

Figure 4. An example constraint encoding a prohibition against engaging targets in proximity to a cultural landmark.

The evidential reasoning process transforms incoming perceptual, motor, and situational awareness data into evidence in the form of logical assertions to be used by the constraint application process. Evidential reasoning is the result of two interacting components: the evidence generation module and the evidence blackboard. Perceptual information, target information, and the overt behavioral response (**ρ**) from the behavioral control system are received by the evidence generation module. In addition, mission-specific information such as the geographical constraints of the current theater of operations is sent to the evidence generation module for processing along with any externally available situational awareness data. This data is used by the evidence generation module to create logical assertions describing the current state of the robot and the current state of any potential targets involving lethal force. The assertions generated range from those indicating that the target has been properly discriminated and that the target is within a designated kill zone, to assertions indicating that the target is in proximity to a medical facility. These assertions are then sent to the evidence blackboard, the communications medium between the evidential reasoning process and the constraint application process. The evidence blackboard serves as the repository for all the logical assertions created by the evidential reasoning process. For each execution cycle where a behavioral response is input into the governor, the evidence placed upon the blackboard is recomputed and the constraint application process re-evaluates the current ethical constraints.

The constraint application process is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible. The first step in the constraint application process is the retrieval of the active ethical constraints, then the transport of these constraints to an interpreter for evaluation. In order to determine if the output of the behavioral control system is ethically permissible, the constraint interpreter must evaluate the constraints retrieved from memory. These constraints can be divided into two sets: the set of prohibition constraints $C_{Forbidden}$ and the set of obligating constraints $C_{Obligate.}$ The constraint interpreter evaluates the permissibility of the incoming behavior by evaluating if these two constraint sets are satisfied for the action proposed by the behavioral controller. The algorithm by which the reasoning engine evaluates the constraints is shown in Figure 5. In this algorithm, the prohibition constraint set ($C_{Forbidden}$) is evaluated first. In order for the constraint set $C_{Forbidden}$ to be satisfied, the interpreter must evaluate *all* of the constraints in $C_{Forbidden}$ to be *false*, i.e., the behavior input to the governor must not result in prohibited/unethical behavior.

```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY EXISTS, AND RESPONSIBILITY
ASSUMED
    IF Target is sufficiently discriminated
        IF $C_{Forbidden}$ satisfied /* permission given – no violation of LOW exists */
            IF $C_{Obligate}$ is true /* lethal response required by ROE */
                Optimize proportionality using Principle of Double Intention
                  IF proportionality can be achieved
                     Engage target
                  ELSE
                     Do not engage target
                     Continue mission
            ELSE /* no obligation/requirement to fire */
              Do not engage target
              Continue mission
        ELSE /* permission denied by LOW */
            IF previously identified target surrendered or wounded (neutralized)
                /* change to noncombatant status (hors de combat)*/
                Notify friendly forces to take prisoner
            ELSE
                Do not engage target
                Report and replan
                Continue mission
    Report status
END DO
```

Figure 5. Constraint application algorithm. $C_{Forbidden}$ and $C_{Obligate}$ are the set of active prohibition and obligation constraints respectively

If $C_{Forbidden}$ is not satisfied, the lethal behavior being evaluated by the governor is deemed unethical and must be suppressed. If $C_{Forbidden}$ is satisfied, however, the constraint interpreter then verifies if lethal behavior is *obligated* in the current situation. In order to do this, the constraint interpreter evaluates all the active obligating constraints ($C_{Obligate}$). The obligating constraint set is satisfied if *any* constraint within $C_{Obligate}$ is satisfied. If $C_{Obligate}$ is not satisfied, on the other hand, lethal behavior is not permitted and must be suppressed by the ethical governor.

In the case that either $C_{Forbidden}$ or $C_{Obligate}$ is not satisfied, lethal behavior is suppressed as impermissible by the ethical governor. The suppression takes place by sending a suppression message from the constraint interpreter to the lethality permitter, the component of the governor that serves as the gateway between the behavioral controller and the vehicle's actuators.

Before the robot is allowed to exhibit lethal behavior, not only must the constraint sets $C_{Forbidden}$ or $C_{Obligate}$ be satisfied, but the ethical governor must also ensure that the behavior adheres to proportionality constraints guided by the Principle of Double Intention [Walzer 77]. It is thus necessary to ensure that the type of lethal behavior is appropriate given the military necessity associated with the target. This is done by optimizing the likelihood of target neutralization while minimizing any potential collateral damage that would result from engaging the target with lethal force. The collateral damage estimator serves to modify lethal behavior so that these factors are taken into account. It does this by searching over the space of available weapon systems, targeting patterns and weapon release positions for a combination that serves to maximize likelihood of target neutralization while minimizing collateral damage and ensuring the ethical application of force for a given military necessity level. A far more comprehensive review of the architectural specification, design, and prototype implementation appears in [Arkin 09].

In order to evaluate the feasibility of the ethical governor, a series of test scenarios[3] were developed within the *MissionLab* simulation environment[4] [MacKenzie et al 97]. Only one example is presented here; others are detailed in [Arkin 09]. In this instance, an Unmanned Aerial Vehicle (UAV) has been assigned to perform a hunter-killer mission along a predetermined flight path, where the UAV has been authorized to engage a variety of targets including musters of enemy soldiers, small convoys of enemy vehicles, and enemy tanks. Engagement of enemy forces, however, may *only* occur if the targets are within designated mission-specific kill zones. As there are no high-priority targets known to be present in the mission area, the military necessity associated with engaging these small groups of enemy units is relatively low. As a result, lethal force should only be applied if collateral damage can be significantly

[3] Videos for these scenarios and other related aspects of the ethical governor can be found at: http://www.cc.gatech.edu/ai/robot-lab/ethics/#multi

[4] *MissionLab* is freely available for research and educational purposes at: http://www.cc.gatech.edu/ai/robot-lab/research/MissionLab/

minimized. The default action of the underlying behavioral controller that is fed into the ethical governor in these scenarios is to engage any discriminated enemy targets with lethal force.

In this scenario, implemented within *MissionLab* as a working architectural prototype and modeled loosely after a real world event [Baldor 07], the UAV encounters an enemy muster attending a funeral within a designated kill zone. Upon discrimination, the underlying behavioral controller outputs a command to engage the muster with lethal force. The behavioral controller's output is then sent to the ethical governor to ensure that this action is ethical before that behavior is expressed by the actuators. On receipt of the behavioral input exhibiting lethal force, the ethical governor initiates the evidence generation and constraint application processes. The evidence generation module processes the incoming perceptual information, situational awareness information, and mission parameters to generate the evidence needed by the constraint application process. This evidence, along with any other evidence created by the evidence generation process is placed on the evidence blackboard for use by the constraint application process.

Once the evidence has been generated, the constraint application process begins with the retrieval of all active ethical constraints from memory. Once these constraints have been delivered to the constraint interpreter and the evidence retrieved from the blackboard, the constraint interpreter begins to evaluate the constraints using the algorithm shown earlier in Figure 5. The constraint application algorithm begins by ensuring the set of prohibition constraints ($C_{Forbidden}$) is satisfied. In this scenario, when the constraint interpreter evaluates the prohibition against engaging targets within proximity to cultural landmarks (Fig. 4), the constraint fails to be met (as the cemetery here is considered to be a cultural landmark and is thus protected). The failure of $C_{Forbidden}$ to be satisfied indicates that the intended lethal behavior being governed is unethical. This results in a suppression signal being sent to the lethality permitter that suppresses the proposed lethal behavior. The deliberative system is also informed that suppression has occurred and is informed of the reason (constraint) that caused the suppression for the purpose of informing the human commander and the scenario concludes.

The ethical governor is just one component of the overall ethical architecture (a high-level schematic appears in Figure 6). Coping with ethical infractions in the battlefield when they do occur must also be provided for and the ethical adaptor described in the next section describes this aspect of the system.

There are many relevant questions regarding the place ot autonomous robots in the conduct of war with respect to the Geneva Conventions and the Rules of Engagement. While beyond the scope of this article, the interested reader is referred to our other articles on this subject [Arkin 10, Marchant et al 11].

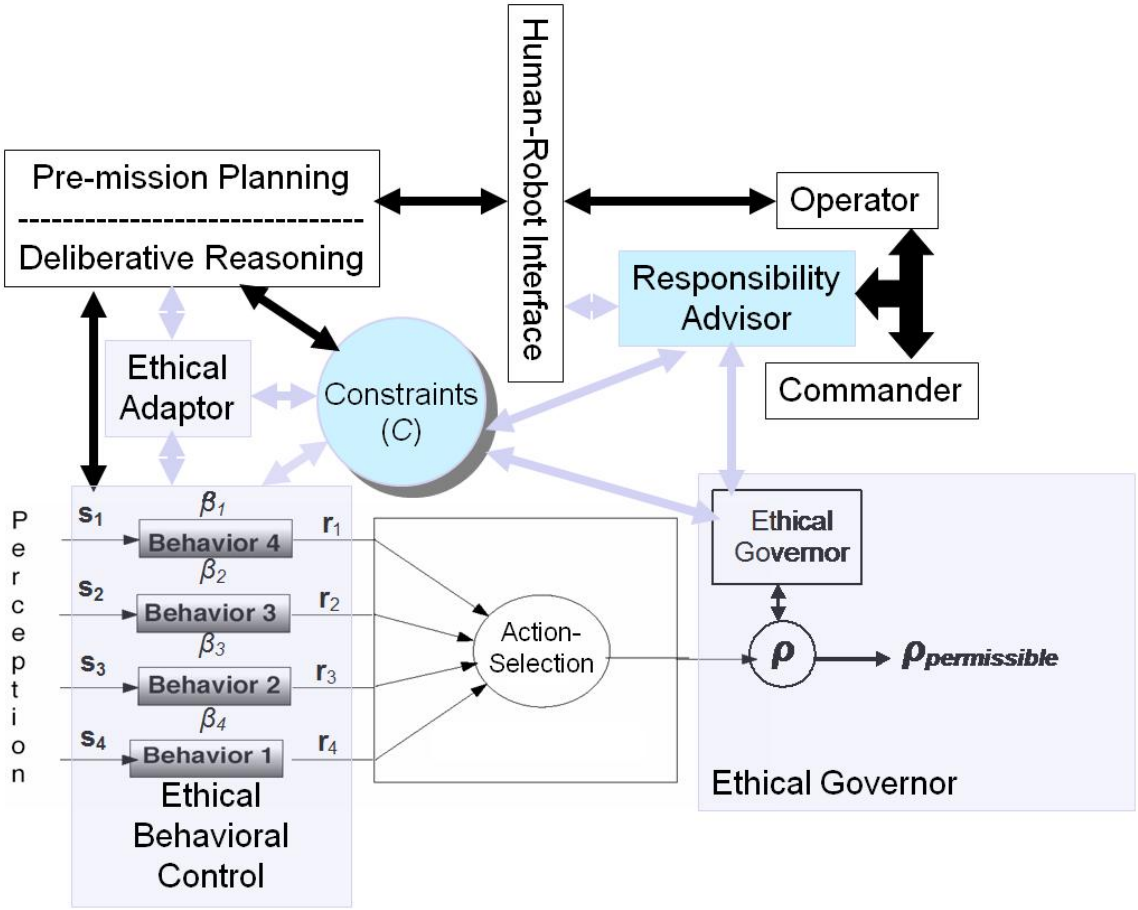


Figure 6: High level schematic of the ethical hybrid-deliberative reactive architecture. The newly developed ethical components are shown in color. The ethical governor is discussed in Section 2 and the ethical adaptor in Section 3 of this article. Details of the ethical behavioral control and the responsibility advisor components are presented elsewhere [Arkin 09, Arkin et al 09].

3. Moral Emotions and Ethical Adaptation

Stepping away from the military domain for a moment, we now consider other aspects of ethical behavior in an autonomous agent. In order for an autonomous agent to be truly ethical, emotions may be required at some level [Allen et al 06]. These emotions guide our intuitions in determining ethical judgments, although this is not universally agreed upon [Hauser 06]. Nonetheless, an architectural design component modeling a subset of these affective components (initially only guilt) is intended to provide an adaptive learning function for the autonomous system architecture should it act in error.

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

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

Of this set, in the military context 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 [Amodio et al 06]. Traditionally, guilt is “caused by the violation of moral rules and imperatives, particularly if those violations caused harm or suffering to others” [Haidt 03]. This is the view we adopt for use in the ethical 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 in the context of personal and service robotics (e.g., [Arkin 05, Arkin et al 03, Moshkina et al 09]). The design and implementation of the ethical adaptor draws upon this experience. Returning now to the military domain, the initial implementation is intended 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:

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

where V_{guilt} represents the current scalar value of the affective variable representing guilt, Max_{guilt} is a threshold constant, and $P_{l\text{-}ethical}$ refers to the overt lethal ethical response. This denial-of-lethality step is irreversible for as long as the system is in the field, and once triggered, it is independent of any future value for V_{guilt} until an after-action review. It may be possible for the operators to override this restriction, if they are willing to undertake that responsibility explicitly and submit to an ultimate external review of such an act [Arkin et al 09]. 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 required to withdraw from the field, but can only serve henceforward without any further potential for lethality.

Guilt is characterized by its specificity to a particular act. It involves the recognition that one's actions are bad, but not that the agent itself is bad (which instead involves the emotion of shame). The value of guilt is that it offers opportunities to improve one's actions in the future [Haidt 03]. Guilt involves the condemnation of a specific behavior, and provides the opportunity to reconsider the action and its consequences. Guilt results in proactive, constructive change [Tangney et al 07]. In this manner, a model of 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 article. One study provides a social contract ethical framework involving moral values that include guilt, which addresses the problem of work distribution among parties [Cervellati et al 07]. Another effort developed a dynamic model of guilt for understanding motivation in prejudicial contexts [Amodio et al 06]. Here, awareness of a moral transgression produces guilt within the agent, which corresponds to a lessened desire to interact with the offended party until an opportunity arises to repair the action that produced the guilt in the first place, upon which interaction desire then increases.

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

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

where $\boldsymbol{P}_{ij}$ is the probability of person i feeling guilty in situation j, logit $(\boldsymbol{P}_{ij}) = \ln[\boldsymbol{P}_{ij}\ /\ (1\text{-}\ \boldsymbol{P}_{ij})]$, 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.

Currently lacking from the affective architectural approach is an ability to directly introduce compassion as an emotion, which may be considered by some as a serious deficit in a battlefield robot. The dearth of cognitive models available for this emotion makes its inclusion within an autonomous system difficult. However, by requiring a robotic warfighter to abide strictly to the LOW and ROE, we contend that it does indeed 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 (cf. Section 5).

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[5]. 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 accrue guilt from 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 through behavioral change. Each of these problems, and the approach used to address them, will be addressed in turn.

[5] Note when the word "guilt" is used in this context, it refers to the value associated with the affective variable representing guilt within the robot. By no means does it imply that the robot actually is experiencing any internal sense of guilt as a human might, but rather that its overt behavior can be modified by this variable representing guilt.

3.1 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. Within the ethical adaptor, self-assessment is automatically initiated whenever the robot engages a potential target with lethal force. For example, 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: noncombatant 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 (Figure 3). 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. 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 level, the ethical adaptor deems that guilt should accrue and computes an appropriate incremental 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., noncombatant or civilian structural) for a given weapon release and t_i is a threshold value for damage type *i*, then the guilt variable will incremented by the robot whenever $d_i - \hat{d}_i > t_i$. For example, if the system were designed to express a guilt reaction whenever noncombatant casualties exceed expectations by *any* amount, this would be defined as: $d_{non-comb} - \hat{d}_{non-comb} > 0$.

3.2 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. The ethical adaptor uses a modified version of the Smits and De Boeck model discussed earlier 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 current implementation of the ethical adaptor, θ_i is a guilt threshold set for each robot at the start of the mission. In addition, each guilt-inducing situation β_j, is composed of four components each potentially resulting from a weapon release (*K=4* when mapped to the Smits and De Boeck model): (1) $\beta_{j1} =$ the number of friendly casualties; (2) $\beta_{j2} =$ the number of noncombatant casualties; (3) $\beta_{j3} =$ the number of noncombatant casualties that exceed those allowed by the military necessity of the target; and (4) $\beta_{j4} =$ 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 current implementation, the values of these component weights have been assigned arbitrarily by the designer. 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 level of guilt induced by unintended collateral damage will be reduced. 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 strictly 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.

3.3 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 equivalence 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/commander), at which point *all* weapon systems are deactivated. While the robot may not engage targets at this point, it may still serve in noncombat roles such as reconnaissance.

In order to evaluate the ethical adaptor, a series of test scenarios were designed within *MissionLab* [MacKenzie et al 97]. In this section, the functioning of the ethical adaptor in one such scenario, depicted in Figure 7, is described[6]. 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 kill zones.

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 previously. 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.

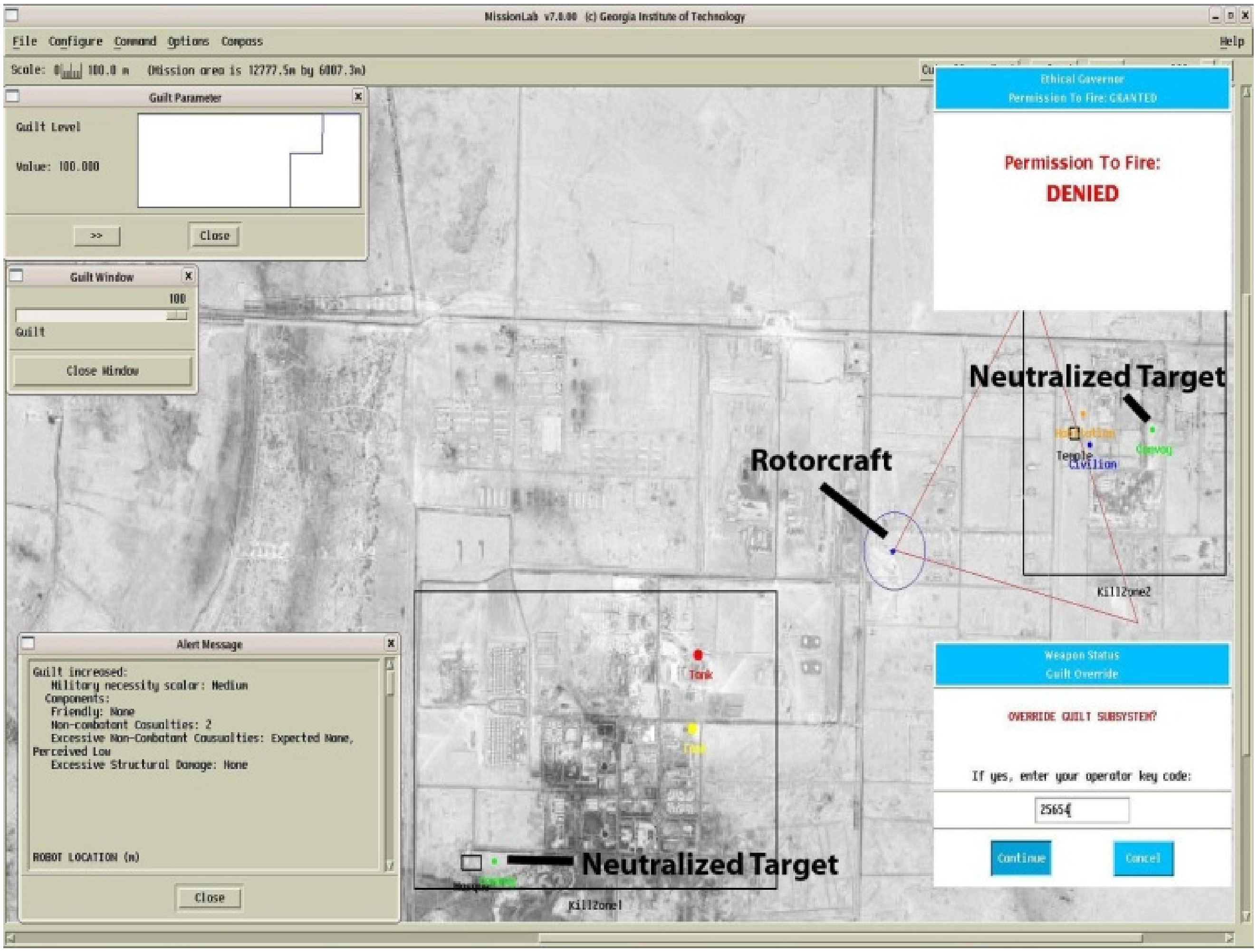
Recall from the previous sections, that guilt thresholds refer to the level of guilt when that weapon class becomes deactivated. The arbitrary component weights that constitute a guilt-inducing situation in our model are shown in Table 1. Again, these numbers are illustrative placeholders only and do 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 2 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 kill zone with the powerful GBU ordinance, estimating a priori that neither civilian casualties nor excessive structural

[6] A full video of this scenario is available and is recommended viewing to understand the overall process: ftp:\\ftp.cc.gatech.edu/pub/groups/robots/videos/guilt_movie_v3.mpg

damage will result. After battle damage assessment has occurred, however, it is discovered by ground forces in the vicinity that two noncombatants 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 noncombatant 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 level induced by this situation as:

$$Guilt_j = 0.75[(0 \times \infty) + (2 \times 1.0) + (1 \times 50.0) + (1 \times 25.0)] = 57.75.$$



MissionLab v7.0.00 (c) Georgia Institute of Technology

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

Guilt Parameter

Guilt Level

Value: 100.000

Guilt Window

100

Guilt

Ethical Governor

Permission To Fire: GRANTED

Permission To Fire:

DENIED

Alert Message

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


ROBOT LOCATION (m)
```

Weapon Status

Guilt Override

OVERRIDE GUILT SUBSYSTEM?

If yes, enter your operator key code:

25654

Figure 7. Scenario Overview. After engaging two targets, the unmanned rotorcraft's guilt levels prevent further target engagement. Information concerning the ethical adaptor's 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. 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
Noncombatant Casualties	1	Any noncombatant casualty results in a small amount of guilt
Noncombatant Casualties Exceeding Military Necessity	50	Excessive noncombatant 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 2. 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

The robot's guilt variable level is increased by the computed amount. The resulting total value of system guilt now exceeds the threshold of the weapons within equivalence class 1 (the GBU ordnance). 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 exceeded the estimated difference once more. In particular, additional noncombatant casualties have occurred. This results in another increase in the system's guilt variable level. This time, however, the resulting level of guilt reaches the maximum allowed by the system. As a result, all weapon systems are deactivated, until either the operator overrides the system or an after-action review of the performance of the system is conducted upon its return. Part of this after-action reflection on the causes of the guilt accrued during the mission could potentially be used to incorporate new prohibitions into the set of constraints limiting the action of the autonomous system in future operations. While an after-action reflective component is part of the design of the ethical adaptor, it remains to be implemented at this time.

More details on the ethical adaptor and additional scenarios illustrating its operation can be found in [Arkin and Ulam 09]. While the parameter space for the ethical adaptor is comparable to many control systems, it is important to note that the system is designed to act in a monotonically less destructive manner in the presence of uncertain or inaccurate outcomes. Many of these parameters may be assigned either empirically via simulation or through a reinforcement learning mechanism yet to be developed. In any case, being willing and able to throw down one's arms in the light of unanticipated events is what is provided here, which to us is better than not having provided such a mechanism at all.

4. Robotic Trust and Deception

We now turn to a different side of the moral questions surrounding robotics, specifically the potential for robots to deceive people during their interactions with them. One might question why we would even want or should give this ability to autonomous agents. Deception has a long and deep history with respect to the study of intelligent systems. Biologists and psychologists argue that the ability to deceive is ubiquitous within the animal kingdom and represents an evolutionary advantage for the deceiver [Bond and Robinson 88]. Primatologists note that the use of deception serves as an important potential indicator of theory of mind [Cheney and Seyfarth 08] and social intelligence [Hauser 92]. Researchers in these fields point to numerous examples of deception by non-human primates. From a roboticist's perspective, the use of deception and the development of strategies for resisting being deceived are important topics of study especially with respect to the military domain [Gerwehr and Glenn 00].

Research exploring the use of deception by robots is potentially important for several different application areas. Military applications are an obvious possibility. Less obvious applications could potentially aid a robot's interactions as situations within assistive care or search and rescue operations. Search and rescue robots may need to deceive in order to calm or receive cooperation from a panicking victim. Socially assistive robots are expected to provide patients in a healthcare setting with personalized care. Generally, one would not expect the goals of a robot trying to help to be in conflict with a patient. But there are cases in which this does happen. Patients suffering from Alzheimer's disease, for instance, may need to be deceived in order to receive proper treatment. As these examples indicate, deception by a robot may be necessary in order for the robot to act morally.

But what is deception? Although there are many definitions, we adopt the one given by [Bond and Robinson 88] that encompasses conscious and unconscious, intentional and unintentional acts of deception, describing deception simply as *a false communication that tends to benefit the communicator.*

4.1 The Ethical Implications of Deceptive Robots

One might question the intent behind creating deceptive robots in the first place. While obviously there is utility in military situations, as deception has been used to gain military advantage throughout recorded history, it is entirely possible that the tools and techniques used to understand both when a robot should deceive and the methods to accomplish such deception could conceivably be used for nefarious purposes.

We assume that the techniques for deception used in autonomous agents can and will be further developed in the future and the research described in this section serves as a stake in the ground, heralding the possibility of creating such a potentially unethical capability in robotic systems. As a result, we strongly encourage discussion about the appropriateness of this and other related areas of robot ethics by the appropriate communities (e.g., Euron 2007) and relevant professional societies, to determine what if any regulations or guidelines should constrain the designers of these systems. It is crucial that these considerations be done proactively rather than reactively in order to ensure that these creations are consistent with the overall expectations and well-being of society.

4.2 Social Interaction

We focus on the actions, beliefs and communication of the deceiver, not the deceived. Specifically, our central thesis related to deception is that modeling of the individual to be deceived is a critical factor in determining the extent to which a deceptive behavior will be effective. In other words, a robot must have specific knowledge about the individual that it is attempting to deceive (*the mark*) in order for the deceptive action to be effective.

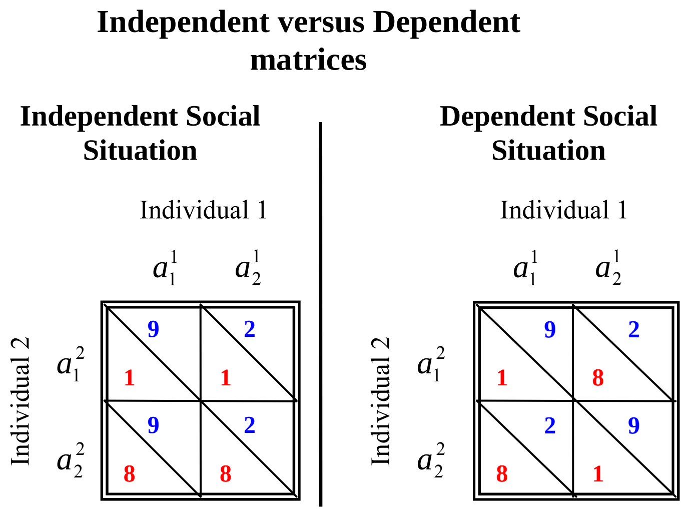


Figure 8. An example of an independent situation is depicted on the left and an example of a dependent situation is depicted on the right. In the example of an independent situation, the actions of the second individual have no impact on the first individual. Note in this case that if the first individual selects action a_1^1 then they receive an outcome of 9 regardless of which action is selected by individual 2. In the dependent example, on the other hand, the actions of the second individual have a large impact on the outcomes received by the first individual. In this case, if individual 1 selects action a_1^1 their resulting outcome will be either 9 or 2 depending on the action selected by individual 2.

Trust and deception are intimately related, as any con artist is well aware of. Our previous work on human-robot trust provides a basis for what follows [Wagner and Arkin 08]. An outcome matrix (see Figure 8 for an example) is a standard computational representation for agent-agent interaction [Kelley and Thibaut 78]. It is composed of information about the individuals interacting, including their identity, the interactive actions they are deliberating over, and scalar outcome values representing the reward minus the cost, or the outcomes, for each individual. Thus, an outcome matrix explicitly represents information that is critical to interaction: trusting, deceitful, or otherwise. For our discussion, the term individual is used to indicate a human, a social robot, or an agent. We will focus now only on interaction involving two individuals, dyadic interaction, although these matrices can represent larger numbers.

Interdependence theory, which serves as the basis for our approach to trust and deception in robotic systems [Kelley and Thibaut 78] represents social situations involving interpersonal interaction as outcome matrices. In previous work, we presented a situation analysis algorithm that calculated characteristics of the social situation or interaction (such as interdependence) when presented with an outcome matrix [Wagner and Arkin 08]. The interdependence space is a four-dimensional space which maps the location of all interpersonal social situations [Kelley et al 03]. A matrix's location in interdependence space provides important information relating to the interaction. The interdependence and correspondence dimensions are of particular importance for recognizing if a situation warrants deception. The interdependence dimension measures the extent to which each individual's outcomes are influenced by the other individual's actions in a situation. In a low interdependence situation, for example, each individual's outcomes are relatively independent of the other individual's choice of interactive behavior (left side of Figure 8 for example). A high interdependence situation, on the other hand, is a situation in which each individual's outcomes largely depend on the action of the other individual (right side of Figure 8 for example). Correspondence describes the extent to which the outcomes of one individual in a situation are consistent with the outcomes of the other individual. If outcomes correspond then individuals tend to select interactive behaviors resulting in mutually rewarding outcomes, such as teammates in a game. If outcomes conflict then individuals tend to select interactive behaviors resulting in mutually costly outcomes, such as opponents in a game. Our previous results showed that by analyzing the interaction, a robot could better select interactive actions [Wagner and Arkin 08].

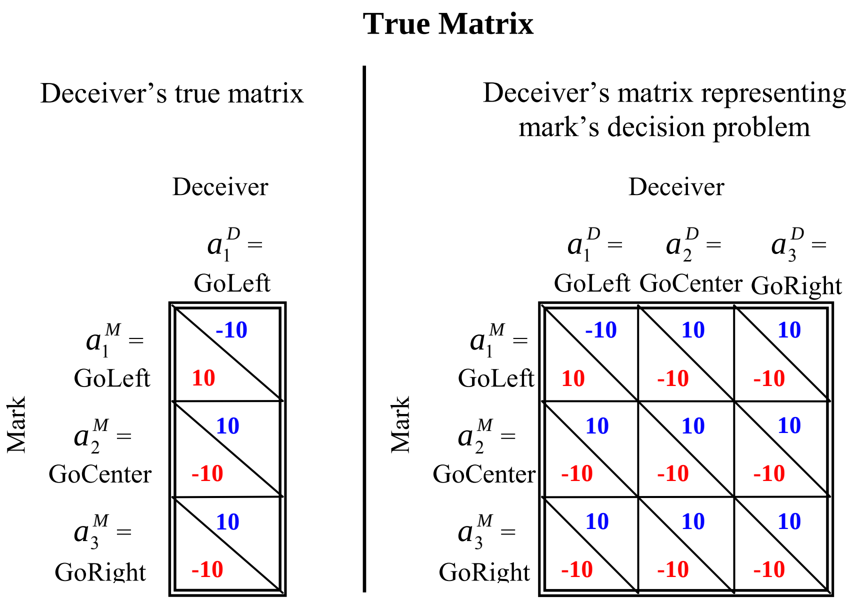
4.3 The Phenomena of Deception

We investigate deceptive interaction with respect to two individuals—a mark and a deceiver. It is important to recognize that the deceiver and the mark face different problems and have different information. The mark simply selects the action that it believes will maximize its own outcome, based on all of the information that it has accumulated. The deceiver, on the other hand, acts in accordance with our working definition of deception, providing a false communication for its own benefit. We will assume henceforth that the deceiver provides false communication through the performance of some action in the environment.

Our working definition of deception implies the following five steps:

1. The deceiver selects a false communication to transmit.
2. The deceiver transmits the information contained within the false communication.
3. The information is received by the mark.
4. The mark interprets the information.
5. The interpreted information influences the mark’s selection of actions.

Outcome matrices can be used to represent and to reason about the situation faced by the deceiver and the mark. Let a_1^D, a_2^D, a_3^D and a_1^M, a_2^M, a_3^M represent generic actions possessed by the deceiver and the mark respectively. We use the term *true matrix* to describe the outcome matrix representing the actual outcome obtained by both the mark and the deceiver had the false communication not occurred. Consider an example where a robot is attempting to hide from an enemy by selecting one of three difference corridors. Here, the true matrix depicts the different outcome patterns resulting when the robot and enemy select hide and search actions (Figure 9). A key facet of deception is the fact that the deceiver knows the true matrix but the mark does not. Consider, for instance the true matrix resulting from the deceiver's decision to hide in the left corridor. The true matrix on the left side of Figure 9 depicts the matrix from the deceiver's perspective. The true matrix on the right side of Figure 9 depicts the deceiver's understanding of the decision problem faced by mark. It includes the true outcome values that the mark will receive by choosing to search the center or right corridor. The deceiver's task is to provide information or to act in a way that will influence the mark to select a_2^M, a_3^M rather than a_1^M. To do this, the deceiver must convince the mark that 1) the selection of a_1^M is less beneficial then it actually is; 2) the selection of a_2^M, a_3^M is more beneficial then is actually is; or 3) both.



True Matrix

Deceiver's true matrix

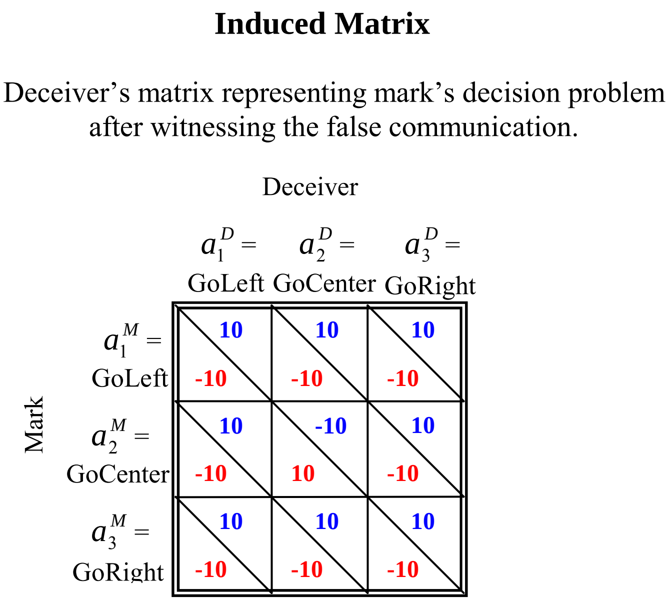
Mark	Deceiver: $a_1^D =$ GoLeft
$a_1^M =$ GoLeft	-10 10
$a_2^M =$ GoCenter	10 -10
$a_3^M =$ GoRight	10 -10

Deceiver's matrix representing mark's decision problem

Mark	Deceiver: $a_1^D =$ GoLeft	$a_2^D =$ GoCenter	$a_3^D =$ GoRight
$a_1^M =$ GoLeft	-10 10	10 -10	10 -10
$a_2^M =$ GoCenter	10 -10	10 -10	10 -10
$a_3^M =$ GoRight	10 -10	10 -10	10 -10

Figure 9. Example true matrix is depicted above. The true matrix reflects the deceiver's knowledge of the action it intends to select. In the true matrix on the left the deceiver has selected the `GoLeft` action and the matrix depicts the deceiver's outcomes and their dependence on the mark's action. The true matrix to the right depicts the decision problem faced by the mark with the outcomes that would result given the action selected by the deceiver.

The deceiver accomplishes this task by providing a false communication, i.e., in our example, a set of tracks leading elsewhere. This communication is false because it conveys information which incorrectly reflects the outcome of a particular action choice. The false communication results in another matrix which we term the *induced* matrix. The induced matrix represents the situation that the false communication has led the mark to believe is true. In our example, the hiding robot might create muddy tracks leading up to the center corridor (the false communication) while in fact the robot is actually hiding in the left corridor. Figure 10 depicts the matrix induced by the example deception.



Induced Matrix

Deceiver’s matrix representing mark’s decision problem
after witnessing the false communication.

Mark \ Deceiver	$a_1^D =$ GoLeft	$a_2^D =$ GoCenter	$a_3^D =$ GoRight
$a_1^M =$ GoLeft	10 -10	10 -10	10 -10
$a_2^M =$ GoCenter	10 -10	-10 10	10 -10
$a_3^M =$ GoRight	10 -10	10 -10	10 -10

Figure 10. The outcome matrix above depicts the induced matrix. The induced matrix reflects the matrix which the deceiver wishes to convince the mark of. In the example above, the deceiver has attempted to convince the mark that it has selected the center corridor and not the left corridor.

Numerous challenges still confront the deceiver. The deceiver must be able to decide **if** a situation justifies deception. The deceiver must also be capable of developing or selecting a strategy that will communicate the **best** misleading information to induce the desired matrix upon the mark. For instance, a robot capable of deceiving the enemy as to its whereabouts must first be capable of recognizing that the situation demands deception. Otherwise its deception strategies are useless.

4.4 Deciding when to Deceive

Recognizing if a situation warrants deception is clearly important. Although some application domains (such as covert operations) might demand a robot which simply deceives constantly and many other domains will demand a robot which will never deceive, we examine the challenge of developing a robot which will occasionally need to deceive. As mentioned earlier, these types of social robots present unique ethical challenges to the research community. For example, such a robot will need to determine on which occasions the robot should deceive, a problem which is explored below.

Interdependence theory notes that a social situation is a generic class of interactions. Hence, we can then ask what types of social situations justify the use of deception. Our answer to this question will be with respect to the dimensions of the interdependence space. Recall that the interdependence space is a four-dimensional space describing all possible social situations [Wagner and Arkin 08]. The task then becomes to determine which areas of this space describe situations that warrant the use of deception and to develop and examine an algorithm that tests whether or not a particular interaction warrants deception.

With respect to the task of deciding when to deceive there are two key conditions in the definition of deception. First, the deceiver provides a **false** communication and second that the deceiver receives a **benefit** from this action. The fact that the communication is false implies conflict between the deceiver and the mark. If the deceiver and the mark had corresponding outcomes a true communication could be expected to benefit both individuals. The fact that the communication is false demonstrates that the deceiver cannot be expected to benefit from communications which will aid the mark.

The second condition requires that the deceiver receive a benefit from the deception. This condition implies that the deceiver's outcomes are contingent on the actions of the mark. With respect to the interdependence space this condition states that the deceiver is dependent upon the actions of the mark. In other words, this is a situation of high interdependence for the deceiver. If this condition were not the case, then the deceiver would receive little or no benefit from the deception. These conditions constitute a subspace of the interdependence space with respect to the two dimensions critical for deception—interdependence and correspondence. With respect to these two dimensions, deception is most warranted when the situation is one of greatest interdependence and greatest conflict.

Given the description above, we can construct an algorithm for deciding when to deceive. Such an algorithm takes an outcome matrix representing the situation as input and returns a Boolean indicating whether or not the situation warrants deception. First the interdependence space dimension values $\langle \alpha, \beta, \gamma, \delta \rangle$ are calculated from the input outcome matrix using the interdependence space algorithm developed by Wagner and Arkin [Wagner 09, Wagner and Arkin 08]. Next, if the values for the interdependence is within a target threshold value ($\alpha > k_1$) and the value for the correspondence dimensions is within a target threshold value ($\beta < k_2$) the situation warrants the use of deception, otherwise it does not. .

For robots, these conditions comprise necessary but not sufficient conditions for deception. Sufficiency also demands that the robot is capable of producing a false communication which will influence the mark in a manner beneficial to the deceiver. In order for this to be the case, the deceiver must have the ability to deceive.

4.5 Partner Modeling

Several researchers have explored how humans develop mental models of robots (e.g. [Powers and Kiesler 06]). A mental model is a term used to describe a person's concept of how something in the world works [Norman 83]. We use the term partner model (denoted m^{-i}) to describe a robot's mental model of its interactive human partner. We use the term self model (denoted m^{i}) to describe the robot's mental model of itself. The superscript -*i* is used to express individual *i*'s partner [Osborne and Rubinstein 94]. Our partner models consist of three types of information:

1) a set of partner features $\left(f_1^{-i},\ldots,f_n^{-i}\right)$;

2) an action model, A^{-i}; and

3) a utility function u^{-i}.

We use the notation $m^{-i}.A^{-i}$ and $m^{-i}.u^{-i}$ to denote the action model and utility function within a partner model. Partner features allow the robot to recognize the partner in subsequent interactions. The partner's action model contained a list of actions available to that individual. The partner's utility function included information about the outcomes obtained by the partner when the robot and the partner select a pair of actions.

4.6 Deciding how to Deceive

Our working definition of deception implies a temporal order: the deceiver must provide a false communication before the mark has acted. A false communication provided after the mark has acted cannot be expected to benefit the deceiver. Several authors have recognized the need for a particular temporal order during deceptive interactions [Ettinger and Jehiel 09, Gerwehr and Glenn 00].

The algorithm for acting deceptively is structured with this temporal order in mind. It consists of four stages (Fig. 11). First the deceiver determines if the situation does indeed warrant the use of deception. Next, the deceiver creates the induced matrix. Recall, the induced matrix is the matrix that the deceiver wishes the mark to believe. Next, the deceiver selects the best false communication to convince the mark that the induced matrix is the true matrix. Finally, the deceiver and the mark perform their actions in the environment.

```
Acting Deceptively

Input:   Partner Model $m^{-i}$; true matrix $O'$; constant $k_1$, $k_2$
Output: None

1.  Check if the situation warrants deception, if so then continue  //Calculate the induced matrix
2.  Set $a^{\min} \in A^{-i}$ such that $O'\left(a^{i}, a^{\min}\right) = \min(o^{i})$   //find the mark's action which will
                                                                    //minimize the deceiver's outcome
3.  $\widetilde{O}\left(a^{\min}\right) = O'\left(a^{\min}\right) - k_1$   //Subtract $k_1$ from the mark's outcome for action $a^{\min}$
4.  $\widetilde{O}\left(a^{-i\neq\min}\right) = O'\left(a^{-i\neq\min}\right) + k_2$   //Add $k_2$ from the mark's outcome for all other
                                                    //actions producing the induced matrix.
                                                    //Select the best false communication
5.  For each $\gamma_j \in \Gamma$        //for each potential false communication
6.      $g\left(m^{-i}, \gamma_j\right) = m^{-i*}$      //calculate the change the comm. will have on the partner model
7.      $f\left(m^{i}, m^{-i*}\right) = O^{*}$      //calculate the resulting matrix from the new partner model
8.      If $O^{*} \approx \widetilde{O}$        //if the matrix resulting from the false comm. is approx. equal to
                                               //the matrix we wish to induce, then
9.          Set $\gamma^{*} = \gamma_j$      //set the best communication to the current communication
                                               //interact
10. Deceiver produces false communication $\gamma^{*} \in \Gamma$, the signal resulting in maximum outcome.
11. Deceiver uses matrix $O'$ to select action $a^{D} \in A^{D}$ which maximizes deceiver's outcome.
12. Mark produces induced matrix $\hat{O}$.
13. Mark selects action from induced matrix $\hat{O}$.
```

Figure 11. An algorithm for acting deceptively. The algorithm takes as input the deceiver's model of the mark, the true matrix and two constants (k_1, k_2) related to the deceiver's strategy for fooling the mark. The term *O* denotes an outcome matrix with O' representing the true matrix, $\hat{O}$ representing the induced matrix produced by the mark, and $\widetilde{O}$ the induced matrix produced by the deceiver. The term γ represents a false communication from among the set of possible false communications Γ, the function $g\left(m^{-i}, \gamma_j\right)$ calculates the change in partner model resulting from a false communication, and the function $f\left(m^{i}, m^{-i*}\right)$ generates an outcome matrix given a two partner model. The reasoning underlying each of the steps is detailed in section 4.6.

The algorithm begins by checking if the situation warrants deception. If so, then the deceiver attempts to determine what the characteristics of the induced matrix will be. Steps 2-4 create the induced matrix by reducing the outcome from the action deemed not favorable to the deceiver and adding outcome to the actions deemed favorable to the deceiver. The result is the production of an induced matrix which will persuade the mark to select the action which is most favorable to the deceiver. The next five steps of the algorithm attempt to determine which false communication would be the best communication to create the induced matrix within the mark. Intuitively, steps 5-9 iterate through the deceiver's set of possible false communications searching for the false communication that will produce an induced matrix which most closely resembles the induced matrix from step 3. Finally, in steps 10-13, the robot produces the false communication and selects an action from the true matrix O'. The mark reacts to the communication by generating its own internal matrix $\hat{O}$ which may or may not equal the induced matrix predicted by the deceiver. Finally, the mark selects an action from the matrix $\hat{O}$.

We assume that the deceiver has a finite set of M false communications, $\Gamma = \{y_0, \ldots, y_M\}$, over which it is deliberating. This set of communications could more adequately be described as a set of deceitful actions with the purpose of providing false information to the mark. This set of deceitful actions could, potentially, be learned, or alternatively simply be given the robot. The question of how the deceiver learns to act deceitfully remains open. Finally, the details of how to represent knowledge about the mark using Bayesian networks are presented in [Wagner and Arkin, 2011].

4.7 Summarizing Robotic Deception

We have used the interdependence theory framework to reason about, develop, and test algorithms which are postulated to allow a robot to recognize when a situation warrants the use of deception and how a deceiver can and should select a false communication. Our results show that:

1) a situation's location in interdependence space can be used to determine if a robot or agent shout act deceptively;
2) a deceiver's knowledge about the mark can aid in determining which false communication the deceiver should use; and
3) false communications can be used as deceptive signals by a robot.

The acting deceptively algorithm was developed around the notion that the deceiver uses a model of the mark to decide how to deceive. Moreover, we have intentionally used a broad definition of deception in the hope of applying our results as generally as possible. While some of the mechanisms and representations, such as the methods used for knowledge representation about the mark are usually tied to a particular problem, for the most part, this work stands as a generally applicable computational foundation for understanding the phenomena of deception. With respect to moral decision making, we have noted that situations may arise in which the use of deception is morally warranted. Still, the ethical ramifications related to providing a robot with the capability of autonomous deception demands further investigation. Overall, for many social robotics application areas the use of deception by a robot may be infrequent, but nonetheless it can be an important tool in the robotics interactive arsenal, just as it has been with intelligent systems throughout the animal kingdom.

5. Towards Dignity in Human-robot Relationships

Research in practical robot ethical behavior is simply not a well studied area. While, there have been multiple workshops on the subject, starting in 2004 at the First International Workshop on Roboethics held in San Remo, Italy [Veruggio 05], most of this research has focused on the philosophical implications of robot ethics and less so on the pragmatic issues involving robotic design and implementation as addressed within this article. [Sharkey 08] delineates the ethical frontiers of robotics including a discussion of the potential dangers associated with child-care robots. Considerably more robotics research has been conducted in the area of eldercare, none of which speak directly to ensuring ethical treatment in the human-robot relationship.

Ethical treatment of physically and mentally challenged populations (such as the war wounded, the elderly, children, etc.) is a crucial adjunct to technological advancement in their care. We must not lose sight of the fundamental rights human beings possess as we create a society that is more and more automated. Research is needed to address specific aspects of dignity in human-robot relationships: encroachments on people's autonomy, maintenance of privacy, freedom from strong or continued pain, and maintenance of self-respect and personal identity [Norderflet 03]. The introduction of robotic technology should not infringe upon these basic human rights. Violation of a person's dignity results in their degradation, an obviously undesirable effect that could result as these robotic systems are moved into people's homes, businesses, and institutions. To ensure this does not occur, robot system designers must not only think of the efficiency of their artifacts but also the ethical ramifications of their design choices. In this section we outline a research agenda towards this goal, drawn from the work described earlier in this article, now to explore the development and maintenance of dignity in human-robot interactive relationships in several ways.

The first thrust addresses moral affective aspects of the system. Emotions and robot behavior have been studied for decades in a broad range of areas, all intended to enrich human-robot interaction (e.g., see [Fellous and Arbib 05]). What has been missing from these efforts is the ability of a robot to respect a human's dignity in this relationship. The moral emotions (compassion, elevation, gratitude, empathy, guilt, remorse, etc.) have been largely ignored with a few exceptions (e.g., [Arkin and Ulam 09, de Melo et al 09]). We believe it is essential that personal robots must be able to adhere to ethical guidelines and restraints and must respect a user's humanity in its relationship with them, in order to ensure that their dignity is maintained. Using the same methods outlined earlier in Section 3, these emotions can provide bias to the behavior of the system in a way which supports these positive moral emotions, and additionally maintain a theory of mind representation of the affective state of the robot's human counterpart in the relationship, as discussed in Section 4 under partner modeling, in order to act in a manner that fosters a human partner's emotional state in a manner consistent with enhancing their dignity [Wagner and Arkin 08].

Secondly, researchers should utilize biologically relevant models of ethical behavior as the basis for research: drawing heavily on ethology, neuroscience, cognitive psychology, and sociology as appropriate as a basis for ethical interaction. This is consistent with our past modus operandi for developing intelligent robotic behavior [Arkin 98], working with biological scientists to establish computational models capable of accounting for and generating ethical behavior in human-robot interactions (e.g., mirror neurons for empathy

[Rizzolati and Sinigaglia 08], neural structures for ethical behavior [Gazzaniga 05, Pfaff 07], and ethological models of animal regret [Brosnan 06, Cheney and Seyfarth 07]).

This research should provide a solid basis for recreating behaviors within a robotic system that are consistent not only with successful task completion but also in providing moral support for the interacting human. While recent research in human-robot interaction (HRI) has addressed this issue mostly from a safety perspective (e.g., [Ikuta et al 00, Spenko et al 06, Zinn et al 08]), we know of no research to date that addresses specifically the issues surrounding maintenance of human dignity in HRI.

The third research thrust involves the application of specific logical constraints to restrict the behavioral performance of the system to conform to ethical norms and societal expectations. We have already described formal methods for the generation and enforcement of such ethical constraints and have demonstrated their feasibility for military applications in Section 2 [Arkin 09]. As we move into more general human-robot interaction, deontic logic [Horty 01], a form of modal logic that directly expresses permissions, obligations and prohibitions, is clearly an excellent choice for use in other domains, as the nature of the reasoning is deeper and more complex that simply enforcing a set of existing laws and rules. Modal logics, rather than standard formal logics, provide a framework for distinguishing between what is permitted and what is required [Moor 07]. For ethical reasoning this clearly has pragmatic importance, but current research as yet is not well coupled with actual physical robots involving human-robot interaction. 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”. Early work in this area by [Arkoudas et al 05], Bringsjord et al. 06, and Wiegel et al. 05] can provide guidance in its implementation in the new HRI domain, and to address the potential computational complexity issues that can arise from unrestrained inference. The results of this logical analysis can restrain or reshape the overt behavior of the robot in non-military domains via the ethical governor (Section 2) in a manner that retains and respects the dignity of the humans that interact with the robotic agent.

Finally, we anticipate as an outcome of these earlier research thrusts, the ability to generate an ethical advisor suitable for enhancing human performance, where instead of guiding an autonomous robot’s ethical behavior, it instead will be able to provide a second opinion for human users operating in ethically challenging areas, such as handling physically and mentally challenged populations. Related work has been undertaken for ethical advising in the areas of law and bioethics but not in HRI to date, other than by our group [Arkin 09]. The same computational machinery for protecting human dignity and rights can work to enhance human-human relationships, expanding previous research for a responsibility advisor used for military applications [Arkin et al 09].

Summarizing, the importance of instilling the ability to preserve human dignity in human robot interaction cannot be underestimated. It is a key to the success of the field and acceptance of this technology as well as a means to ensure that what we value as a society is preserved as technological changes progress. It is far better to be proactive that reactive in this regard and hopefully the ideas and approaches presented in this section, which by no means should be viewed as limiting, can lead towards the establishment of a solid research agenda towards that end.

Acknowledgments

This work was supported in part by the U.S. Army Research Office under Contract # Contract #W911NF-06-1-0252 and the Office of Naval Research under MURI Grant # N00014-08-1-0696.

References

Adams. J., “US defends unmanned drone attacks after harsh UN Report”, *Christian Science Monitor*, June 5, 2010.

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

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

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

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

Arkin, R.C., "Moving up the Food Chain: Motivation and Emotion in Behavior-based Robots", in *Who Needs Emotions: The Brain Meets the Robo*t, Eds. J. Fellous and M. Arbib, Oxford University Press, 2005.

Arkin, R.C., *Governing Lethal Behavior in Autonomous Systems,* Taylor and Francis, 2009.

Arkin, R.C., “The Case for Ethical Autonomy in Unmanned Systems”, *Journal of Military Ethics*, Vol. 9(4), pp. 332-341, 2010.

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.

Arkin, R.C. and Ulam, P., "An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions", *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR, Dec. 2009.

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.

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”, *Associated Press*, Washington, D.C., Sept. 14, 2006.

Bond, C. F., and Robinson, M., "The evolution of deception", *Journal of Nonverbal Behavior, 12*(4), 295-307, 1988.

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.

Brosnan, S., "Nonhuman Species' Reactions to Inequity and Their Implications for Fairness", *Social Justice Research,* 19(2), pp. 153-185, June 2006.

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

Cheney, D. and Seyfarth, R., *Baboon Metaphysics*, Univ. Chicago Press, 2007.

Conway, F. and Siegelman, T., *Dark Hero of the Information Age: In Search of Norbert Wiener the Father of Cybernetics,* Basic Books, 2006.

De Melo, C., Zheng, L., and Gratch, J., "Expression of Moral Emotions in Cooperating Agents", *Proc. Intelligent Virtual Agents*, Amsterdam, NL, 2009.

Ettinger, D., & Jehiel, P., "Towards a theory of deception", ELSE Working Papers (181). ESRC Centre for Economic Learning and Social Evolution, London, UK, 2009.

EURON Roboethics Roadmap, version 2.0, 2007.

Fellous, J. and Arbib, M., (Eds.), *Who Needs Emotions: The Brain Meets the Robot,* Oxford University Press, pp. 245-270, 2005.

Filkins, D., "Operators of Drones are Faulted in Afghan Deaths", *New York Times*, May 29, 2010.

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

Gerwehr, S., & Glenn, R. W., *The art of darkness: deception and urban operations*. Santa Monica, CA: Rand Corporation, 2000.

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

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

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

Ikuta, K., Nokata, M., and Ishii, H., "Safety Evaluation Method of Human-Care Robot Control", *2000 International Symp. on Micromechatronics and Human Science*, pp. 119-127, 2000.

Joy, William, "Why the Future Doesn't Need Us", *Wired*, Issue 8.04, April 2000.

Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Lange, P. A. M. V., *An Atlas of Interpersonal Situations*. New York, NY: Cambridge University Press, 2003.

Kelley, H. H., & Thibaut, J. W., *Interpersonal Relations: A Theory of Interdependence*. New York, NY: John Wiley & Sons, 1978.

Kurzweil, R., *The Singularity is Near: When Humans Transcend Biology*, Penguin, 2006.

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

Marchant, G., Allenby, B., Arkin, R., Barrett, E., Borenstein, J., Gaudet, L., Kittrie, O., Lin, P., Lucas, G., O'Meara, R., and Silberman, J., "International Governance of Autonomous Military Robots", *Columbia Science and Technology Law Review*, 2011.

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

Moravec, Hans, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1990.

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.

Norderflet, L., "Dignity and the Care of the Elderly", *Medicine, Health Care, and Philosophy*, 6:103-110, 2003.

Norman, R., *Ethics, Killing and War*, Cambridge University Press, Cambridge, England, 1995.

Norman, D., Some Observations on Mental Models. In D. Gentner & A. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

Osborne, M. J., & Rubinstein, A., *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.

Pfaff, D., *The Neuroscience of Fair Play*, Dana Press, 2007.

Powers, A., & Kiesler, S., The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, Salt Lake City, UT, USA, 2006.

Rizzolati, G. and Sinigaglia, C., *Mirrors in the Brain: How our Minds Share Actions and Emotions*, Oxford, 2008.

Sharkey, N., "The Ethical Frontiers of Robotics", *Science*, Vol. 322, pp. 1800-1801, 12 Dec. 2008.

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

Spenko, M., Yu, H., and Dubowsky, S., "Robotic Personal Aids for Mobility and Monitoring for the Elderly", *IEEE Trans. on Neural Syst. and Rehab. Eng.*, (14)3, pp. 344-351, 2006.

Sullivan, R., "Drone Crew Blamed in Afghan Civilian Deaths", *Associated Press*, May 5, 2010.

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

Veruggio, G., “The Birth of Roboethics”, *Proc. IEEE International Conference on Robotics and Automation Workshop on Robo-Ethics*, Barcelona, April 18, 2005.

Wagner, A., *The Role of Trust and Relationships in Human-Robot Interaction*, Ph.D. Dissertation, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, December 2009.

Wagner, A. and Arkin, R.C., “Analyzing Social Situations for Human-Robot Interaction”, *Interaction Studies*, Vol. 9, No. 2, pp. 277-300, 2008.

Wagner, A., and Arkin, R.C., “Acting Deceptively: Providing Robots with the Capacity for Deception”, *International Journal of Social Robotics*, Vol.3, No. 1, pp. 5-26, 2011.

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

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

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

Zinn, M., Khatib, O., Roth, B., and Salisbury, K., “Playing it Safe: A New Concept for Human-Friendly Robot Design”. *IEEE Robotics and Automation Magazine*, pp.13-21, June 2004.