Responsibility and Lethality for Unmanned Systems: Ethical Pre-mission Responsibility Advisement

Ronald C. Arkin
Georgia Institute of Technology
85 Fifth Street NW
Atlanta, GA 30308
1-404-894-8209
arkin@cc.gatech.edu

Alan R. Wagner
Georgia Institute of Technology
85 Fifth Street NW
Atlanta, GA 30308
1-404-894-9311
alan.wagner@cc.gatech.edu

Brittany Duncan
Georgia Institute of Technology
85 Fifth Street NW
Atlanta, GA 30308
1-404-894-9311
gth656@gmail.gatech.edu

ABSTRACT

This paper provides an overview, rationale, design, and prototype implementation of a responsibility advisor for use in autonomous systems capable of lethal target engagement. The ramifications surrounding the potential use of operator overrides is also presented. The results of this research have been integrated into the MissionLab mission specification and demonstrated on a relevant military scenario.

1. INTRODUCTION

The advent of autonomous lethal robotic systems is well underway and it is a simple matter of time before autonomous engagements of targets are present on the battlefield. We have written extensively on this subject [1-3]. This article focuses specifically on the issue of ethical responsibility for the use of such systems and how an automated human-robot advisor can assist in making informed decisions prior to the deployment of robotic weaponry.

This is obviously not without controversy. Sparrow [4] argues that any use of "fully autonomous" robots is unethical due to the *Jus in Bello* requirement that someone must be responsible for a possible war crime. He argues that while responsibility could ultimately vest in the commanding officer for the system's use, it would be unjust to both that individual and any resulting casualties in the event of a violation. Nonetheless, due to the increasing tempo of warfare, he shares our opinion that the eventual deployment of systems with ever increasing autonomy is inevitable. We agree that it is necessary that responsibility for the use of these systems must be made clear, but do not agree that it is infeasible to do so.

Several existing weapons systems are already in use that deploy lethal force autonomously to some degree (e.g., land mines, cruise missiles, phalanx system) and they (with the exception of anti-personnel land mines, due to their lack of discrimination, not responsibility attribution) are not generally considered to be unethical. He also neglects to consider the possibility of the embedding of prescriptive ethical codes within the robot itself, which can govern its actions in a manner consistent with the Laws of War (LOW) and Rules of Engagement (ROE), thus weakening his claim. While Sparrow is "quite happy to allow that robots will become capable of increasingly sophisticated behavior in the future and perhaps even of distinguishing between war crimes and legitimate use of military force", the underlying question regarding responsibility, he contends, is not solvable. It is our belief, however, that by making the assignment of responsibility transparent and explicit, through the use of a responsibility advisor at all steps in the deployment of these systems, that this problem is solvable.

Asaro [5] similarly argues from a position of loss of attribution of responsibility, but broaches the subject of robots possessing "moral intelligence". His definition of a moral agent seems applicable, where an agent adheres to a system of ethics, which it employs in choosing the actions that it takes or refrain from taking. He also considers legal responsibility, which he states will compel roboticists to build ethical systems in the future. He notes, similar to what is proposed here, that if an existing set of ethical policy (e.g., LOW and ROE) is replicated by the robot's behavior, it enforces a particular morality through the robot itself. It is in this sense we strive to create such an ethical architectural component for unmanned systems, where that "particular morality" is derived from International Conventions.

One of the earliest arguments encountered based upon the difficulty to attribute responsibility and liability to autonomous agents in the battlefield was presaged by Perri [6]. He assumes "at the very least the rules of engagement for the particular conflict have been programmed into the machines, and that only in certain types of emergencies are the machines expected to set aside these rules". We personally do not trust the view of setting aside the rules by the autonomous agent itself, as it begs the question of responsibility if it does so, but it may be possible for a human to assume responsibility for such deviation if it is ever deemed appropriate (and ethical) to do so. While he rightly notes the inherent difficulty in attributing responsibility to the programmer, designer, soldier, commander, or politician for the potential of war crimes by these systems, it is believed that a deliberate assumption of responsibility by human agents for these systems can at least help focus such an assignment when required. A central part of the architecture in this article is a responsibility advisor, which specifically addresses these issues, although it would be naïve to say it will solve all of them. Often assigning and establishing responsibility for human war crimes, even through International Courts, is quite daunting.

The elimination of the need for an autonomous agent's claim of self-defense as an exculpation of responsibility through either justification or excuse is of related interest, which is a common occurrence during the occasioning of civilian casualties by human soldiers [7]. Robotic systems need make no appeal to self-defense or self-preservation in this regard, and thus can and should value civilian lives above their own continued existence. Of course there is no guarantee that a lethal autonomous system would be given that capability, but to be ethical we would contend that it must. This is a condition that a human soldier likely could not easily or ever attain to, and as such it would allow an ethical autonomous agent to potentially outperform a human in this regard. This is discussed at length in [8]. The system's use of lethal force does not preclude collateral damage to civilians and their property during the conduct of a military mission according to the Just War Principle of Double Effect[1], only that no claim of self-defense could be used to justify any such incidental deaths. It also does not negate the possibility of the autonomous system acting to defend fellow human soldiers under attack in the battlefield.

The overall architecture for this lethal ethical autonomous system is described in [8] and is depicted in Figure 1. The architectural design must implement these processes effectively, efficiently, and be consistent with the constraints derived from the LOW and ROE. This article focuses solely on the responsibility advisor that forms a part of the human-robot interaction component used for pre-mission planning and managing operator overrides. It advises in advance of the mission, the operator(s) and commander(s) of their ethical responsibilities should the lethal autonomous system be deployed for a specific battlefield situation. It requires their explicit acceptance (authorization) prior to its use. It also informs them regarding any changes in the system configuration, especially in regards to the constraint set that encodes the LOW and ROE. In addition, it requires operator responsibility acceptance in the event of a deliberate override of an ethical constraint preventing the autonomous agent from acting.

Colin et al [9] note that "as systems get more sophisticated and their ability to function autonomously in different context and environment expands, it will become important for them to have 'ethical subroutines' of their own… these machines must be self-governing, capable of assessing the ethical acceptability of the options they face" The architectural approach advocated in this architecture embodies that spirit, but is considerably more complex than simple subroutines.

2. RESPONSIBILTY ADVISEMENT

A crucial design criterion and associated design component, the **Responsibility Advisor**, must make clear and explicit as best as possible, just where *responsibility* vests, should: (1) an unethical action be undertaken by the autonomous robot as a result of an operator/commander override; or (2) the robot performs an unintended unethical act due to some representational deficiency in the constraint set or in its application either by the operator or within the architecture itself. To do so requires not only suitable training of operators and officers as well as appropriate architectural design, but also an on-line system that generates awareness to soldiers and commanders alike about the consequences of the deployment of a lethal autonomous system. It must be capable of providing reasonable explanations for its actions regarding lethality, including refusals to act.

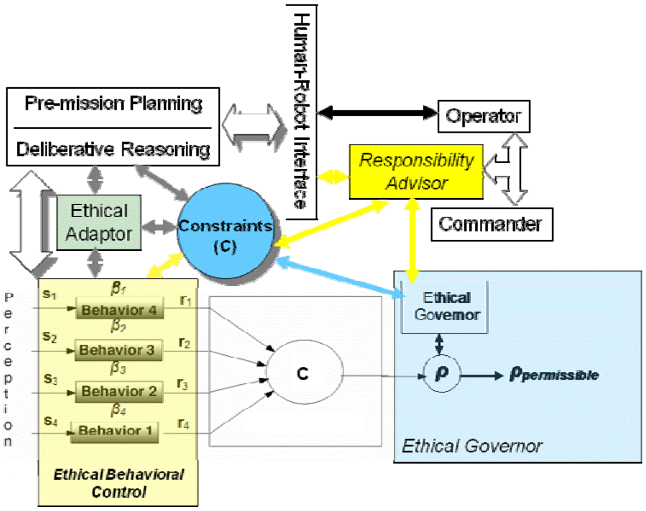


Figure 1: Ethical Architecture (See [8] for details)

Certainly the agent should never intend to conduct a forbidden lethal action, and although an action may be permissible, it should also be deemed obligatory in the context of the mission (military necessity) to determine whether or not it should be undertaken. So in this sense, we argue that any lethal action undertaken by an unmanned system must be obligatory and not solely permissible, where the mission ROE define the situation-specific lethal obligations of the agent and the LOW define absolutely forbidden lethal actions. Although it is conceivable that permissibility alone for the use of lethality is adequate, we will require the provision of additional mission constraints explicitly informing the system regarding target requirements (e.g., as part of the ROE) to define exactly what constitutes an acceptable action in a given mission context. This assists with the assignment of responsibility for the use of lethality. Laws of War and related ROE determine what are absolutely forbidden lethal actions; and Rules of Engagement mission requirements determine what is obligatory lethal action, i.e., where and when the agent must exercise lethal force. Permissibility alone is inadequate.

"If there are recognizable war crimes, there must be recognizable criminals" [10]. The theory of justice argues that there must be a trail back to the responsible parties for such events. While this trail may not be easy to follow under the best of circumstances, we need to ensure that accountability is built into the ethical architecture of an autonomous system to support such needs. On a related note, does a lethal autonomous agent have a right, even a responsibility, to refuse an unethical order? The answer is an unequivocal yes. "Members of the armed forces are bound to obey only lawful orders" [11]. What if the agent is incapable of understanding the ethical consequences of an order, which indeed may be the case for an autonomous robot? That is also spoken to in military doctrine: It is a defense to any offense that the accused was acting pursuant to orders unless the accused knew the orders to be unlawful or a person of ordinary sense and understanding would have known the orders to be unlawful [12].

That does not absolve the guilt from the party that issued the order in the first place. During the Nuremberg trials it was not sufficient for a soldier to merely show that he was following orders to absolve him from personal responsibility for his actions.

[1] The Principle of Double Effect, derived from the Middle Ages, asserts that as long as 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).

Two other conditions had to be met [13]: (1) The soldier had to believe the action to be morally and legally permissible; and (2) The soldier had to believe the action was the only morally reasonable action available in the circumstances. For an ethical robot it should be fairly easy to satisfy and demonstrate that these conditions hold due to the closed world assumption, i.e., the robot's beliefs can be well-known and characterized, and perhaps even inspected (assuming the existence of explicit representations and not including learning robots in this discussion). Thus the responsibility returns to those who designed, deployed, and commanded the autonomous agent to act, as they are those who controlled its beliefs.

Matthias [14] speaks to the difficulty in ascribing responsibility to an operator of a machine that employs learning algorithms since the operator is no longer in principle capable of predicting the future behavior of that agent any longer. The use of subsymbolic machine learning is not currently advocated at this time for any of the ethical architectural components. We accept the use of inspectable changes by the lone adaptive component used within the ethical components of the architecture, (i.e., the ethical adaptor). This involves change in the explicit set of constraints that governs the system's ethical performance. Matthias notes "as long as there is a symbolic representation of facts and rules involved, we can always check the stored information and, should this be necessary, correct it." We contend that by explicitly informing and explaining to the operator of any changes made to the ethical constraint set by the reflective activities of the ethical adaptor prior to the agent's deployment on a new mission, and ensuring that any changes due to learning do not occur during the execution of a mission, an informed decision by the operator can be made as to the system's responsible use. Matthias concludes that "if we want to avoid the injustice of holding men responsible for actions of machines over which they could not have sufficient control, we must find a way to address the responsibility gap in moral practice and legislation." In any case, the responsibility advisor is intended to make explicit to the operator of an ethical agent the responsibilities and choices he/she is confronted with when deploying autonomous systems capable of lethality.

Responsibility acceptance occurs at multiple levels within the architecture:

1. Command authorization of the system for a particular mission.
2. Override responsibility acceptance.
3. Authoring of the constraint set that provides the basis for implementing the LOW and ROE, which entails responsibility – both from the ROE author and by the diligent translation by a second party into a machine recognizable format. It should be noted that failures in the accurate description, language, or conveyance of the ROE to a soldier have often been responsible or partially responsible for the unnecessary deaths of soldiers or violations of the LOW [15]. Great responsibility will vest in those who both formulate the ROEs for lethal autonomous systems to obey, and similarly for those who translate them into machine usable forms. Mechanisms for verification, validation, and testing must be an appropriate part of any plan to deploy such systems.
4. Verification that only military personnel are in charge of the system. Only military personnel (not civilian trained operators) have the legal authority to conduct lethal operations in the battlefield.

The remainder of this section focuses primarily on two aspects of responsibility assignment: authorizing a lethal autonomous system for a mission, and the use of operator controlled overrides.

2.1 Command Authorization

Obligating constraints provide the sole justification for the use of lethal force within the ethical autonomous agent. Forbidding constraints prevent inappropriate use, so the operator must be aware of both, but in particular, responsibility for any mission-specific obligating constraints that authorize the use of lethality must be acknowledged prior to deployment. Klein [16] identifies several ways in which accountability can be maintained for armed UVs:

1. "Kill Box" operations, where a geographic area is designated where the system can release its weapons after proper identification and weapon release authority is obtained.
2. Targets located and identified prior to an unmanned vehicle (UV) arriving on scene. Once on scene, the UV receives target location and a "clear to fire" authorization.
3. "Command by Negation", where a human overseer has responsibility to monitor targeting and engagements of a UV but can override the automated weapons systems.

Our approach within the ethical architecture differs in several respects. Kill box locations must be confirmed in advance of the mission as part of the ROE and are encoded as constraints. Candidate targets and target classes must be identified in advance, and must also be confirmed by the system during the operation itself prior to engagement. Permission-to-fire is granted during the mission in real-time if obligating constraints require, not simply upon arrival at the scene. This use of obligatory constraints, derived from the ROE, assists in the acceptance of responsibility for the use of lethal action by the operator, due to transparency regarding what the system is permitted to achieve with lethal force. To establish this responsibility, prior to deployment the operator must acquire and acknowledge possessing an explicit understanding of the underlying constraints that determine how lethality is governed in the system. In addition to advanced operator training, this requires making clear, in understandable language, exactly which obligations the system maintains regarding its use of lethal force for the given mission and specifically what each one means. These explanations must clearly demonstrate that:

- Military necessity is present and how it is established
- How combatant/target status is determined
- How proportional response will be determined relative to a given threat

The operator is required to visually inspect every single obligating constraint in the architecture's short-term memory (STM) prior to mission deployment, understand its justification, and then acknowledge its use. This results in responsibility acceptance. The user interface must facilitate and support this operation. The implications of LOW and ROE-derived constraints that reside in long-term memory (LTM) must be conveyed to the operator earlier through qualification training for use of the system in the field prior to actual deployment. Any changes in LTM constraint representations that occur after training must be communicated to the operator in advance of use, and acknowledgment of the understanding of the consequences of these changes accepted in writing.

The results of previous experience and/or the consultations of expert ethicists regarding similar previous mission scenarios can also be presented to the operator for review. This can help ensure that mistakes of the past are not repeated, and that judgments from ethical experts are included in the operator's decision whether or not to use the lethal autonomous system in the current context, effectively providing a second or third opinion prior to use.

2.2 Design for Mission Command Authorization

Several architectural design features are necessary for mission authorization. They involve a method to display the mission's active obligating constraints and to allow the operator to probe to whatever depth is required in order to gain a full understanding of the implications of their use, including expert opinion if requested. This interface must:

1. Require acknowledgment that the operator has been properly trained for the use of an autonomous system capable of lethal force, and understands all of the forbidding constraints in effect as a result of their training. It must also confirm the date of their training and if any updates to forbidding constraints in LTM have occurred since then to ensure he/she is aware of and accepts them.
2. Present all obligations authorizing the use of lethal force by providing clear explanatory text and justification for their use at multiple levels of abstraction. The operator must accept them one by one via a checkbox in order to authorize the mission.
3. Recall previously stored missions (both human and autonomous) and their adjudged ethical appropriateness, as obtained from expert ethicists. This may require additional operator input concerning the location, type, and other factors regarding the current mission, beyond the existing ROE constraint set. These results must be presented in a clear and unambiguous fashion, and the operator must acknowledge having read and considered these opinions.
4. A final authorization for deployment must be obtained.

The system is now ready to conduct its mission, with the operator explicitly accepting responsibility for his role in committing the system to the battlefield.

2.3 The Use of Ethical Overrides

Walzer [10] recognizes four distinct cases regarding the Laws of War and the theory of aggression:

1. LOW are ignored under the "pressure of a utilitarian argument."
2. A slow erosion of the LOW due to "the moral urgency of the cause" occurs, where the enemies' rights are devalued and the friendly forces' rights are enhanced.
3. LOW is strictly respected whatever the consequences.
4. The LOW is overridden, but only in the face of an "imminent catastrophe."

It is our contention that autonomous robotic systems should adhere to case 3, but potentially allow for case 4, where only humans are involved in the override. By purposely designing the autonomous system to strictly adhere to the LOW, this helps scope responsibility, in the event of an immoral action by the agent. Regarding the possibility of overriding the fundamental human rights afforded by the Laws of War, Walzer notes:

"These rights, I shall argue, cannot be eroded or undercut; nothing diminishes them, they are still standing at the very moment they are overridden: that is why they have to be overridden.… The soldier or statesman who does so must be prepared to accept the moral consequences and the burden of guilt that his action entails. At the same time, it may well be that he has no choice but to break the rules: he confronts at last what can meaningfully be called necessity."[10]

This ability and resulting responsibility for committing an override of a fundamental legal and ethical limit should not be vested in the autonomous system itself. Instead it is the province of a human commander or statesman, where they must be duly warned of the consequences of their action by the autonomous agent that is so instructed. Nonetheless, a provision for such an override mechanism of the Laws of War may perhaps be appropriate in the design of a lethal autonomous system, at least according to our reading of Walzer, but should not be easily invoked and must require multiple confirmations in the chain of command before the robot is unleashed from its constraints.

In effect, the issuance of a command override changes the status of the machine from an autonomous robot to that of a robot serving as an extension of the warfighter, and in so doing the operator(s) must accept all responsibility for their actions. These are defined as follows [17]:

- Robot acting as an extension of a human soldier: a robot under the direct authority of a human, especially regarding the use of lethal force.
- Autonomous robot: a robot that does not require direct human involvement, except for high-level mission tasking**;** such a robot can make its own decisions consistent with its mission without requiring direct human authorization, especially regarding the use of lethal force.

If overrides are to be permitted, they must use a variant of the two-key safety precept, (DSP-15 from [18]) but slightly modified for overrides: The overriding of ethical control of autonomous lethal weapon systems shall require a minimum of two independent and unique validated messages in the proper sequence from two different authorized command entities, each of which shall be generated as a consequence of separate authorized entity action. Neither message should originate within the unmanned system launching platform.

The management and validation of this precept is a function of the responsibility advisor. If an override is accepted, the system must generate a message logging this event and transmit it to legal counsel, both within the U.S. military and to International Authorities. Certainly this will assist in making the decision to override the LOW a well-considered one by an operator, simply by the potential consequences of conveying immediately to the powers-that-be news of the use of potentially illegal force. This operator knowledge enhances responsibility acceptance for the use of lethal force, especially when unauthorized by the ethical architecture.

The ethical architecture serves as a safety mechanism for the use of lethal force. If it is removed for whatever reason, the operator must be advised of the consequences of such an act. The system should still monitor and expose any ethical constraints that are being violated within the architecture to the operator even when overridden, if it is decided to use lethality via this system bypass. The autonomous system can still advise the operator of any ethical constraint violations even if the operator is in direct control (i.e., by setting the Permission-To-Fire variable to TRUE, enabling the weapons systems). If such ethical violations exist at the time of weapons deployment, a "two-trigger" pull is advised, as enforced by the autonomous system. A warning from the system should first appear that succinctly advises the operator of any perceived violations, and then and only then should the operator be allowed to fire, once again confirming responsibility for their action by so doing. These warnings can be derived directly from the forbidden constraints while also, if appropriate, providing a warning that there is no obligation to fire under the current mission conditions, i.e., there exists no obligating constraint that is TRUE at the time.

When these constraints are added, either in LTM or STM, the developer must assume responsibility for the formulation of that constraint and its ethical appropriateness before it can be used within a fielded system. Normally this would occur through a rigorous verification and validation process prior to deployment. The basic research conducted in this effort, is intended to be proof of concept only, and will not necessarily create constraints that completely capture the requirements of the battlefield or are intended in their current form for that purpose.

2.4 Design for Overriding Ethical Control

Overriding means changing the system's ability to use lethal force, either by allowing it when it was forbidden by the ethical controller, or by denying it when it has been enabled. As stated earlier, overriding the forbidding ethical constraints of the autonomous system should only be done with utmost certainty on the part of the operator. To do so at runtime should require a direct "two-key" mechanism, with coded authorization by two separate individuals, ideally the operator and his immediate superior. This operation is generally not recommended and, indeed it may be wise to omit it entirely from the design to ensure that operators do not have the opportunity to violate the Laws of War. In this way the system can only err on the side of not firing. The inverse situation, denying the system the ability to fire, does not require a two-key test, and can be done directly from the operator console. This is more of an emergency stop scenario, should the system be prepared to engage a target that the operator deems inappropriate for whatever reasons.

The functional equivalent of an override is the negation of the PTF (Permission-To-Fire) variable that is normally directly controlled by the ethical architecture. This override action allows the weapons systems to be fired even if it is not obligated to do so (F → T) potentially leading to atrocities, or eliminating its obligated right to fire if the operator thinks it is acting in error (T →F). This is accomplished through the use of the exclusive OR function. Table 1 captures these relationships.

Table 1. Logical values for the Permission-to-fire (PTF) variable.

	Governor PTF Setting	Operator Override	Final PTF Value	Comment
1	F (do not fire)	F (no override)	F (cannot fire)	System does not fire as it is not overridden
2	F (do not fire)	T (override)	T (can fire)	Operator commands system to fire despite contrary ethical recommendations
3	T (perm. to fire)	F (no override)	T (can fire)	System is obligated to fire
4	T (perm. to fire)	T (override)	F (cannot fire)	Operator negates system's permission to fire

In case 2, using a graphical user interface (GUI), the operator must be advised and presented with the forbidden constraints he/she is potentially violating. Each violated constraint is presented to the operator with an accompanying text explanation for the reasoning behind the perceived violation and any relevant expert case opinion that may be available. This explanation process may proceed, at the operator's discretion, down to a restatement of the relevant Laws of War if requested. The operator must then acknowledge understanding each constraint violation and explicitly check each one off prior to an override for that particular constraint being rescinded. One or more constraints may be removed by the operator at their discretion. After the override is granted, automated notification of the override is sent immediately to higher authorities for subsequent review of its appropriateness.

Similarly in case 4, the operator must be advised and presented with the ROE obligations she/he is neglecting during the override. One or all of these obligating constraints may be removed. As case 4 concerns preventing the use of force by the autonomous system, the operator can be granted instantaneous authority to set the Permission-to-Fire value to FALSE, without requiring a prior explanation process, a form of emergency stop for weapon release.

3. IMPLEMENTING THE RESPONSIBILTY ADVISOR

The responsibility advisor is currently partially implemented as part of the *MissionLab* behavior specification system [19]. *MissionLab* provides automated pre-mission planning functions to specify detailed mission objectives for the operator to utilize. The user interacts through a design interface tool (the configuration editor) that permits the visualization of a mission specification as it is created. The responsibility advisor serves as a gatekeeper to the mission specification system, preventing unauthorized mission creation as well as counseling users regarding the mission's obligations and prohibitions. The operation of the pre-mission responsibility advisor occurs in five steps:

1. Determine if the user is authorized to conduct the mission and when they were trained. If they are authorized, then:
2. The user selects a mission.
3. The user is presented with plain text descriptions of their mission obligations including related information. If they accept these obligations, then:
4. Present the user with a plain text description of any prohibitions that have changed since their training including related information. If they accept these prohibitions, then:
5. Present final authorization for the mission.

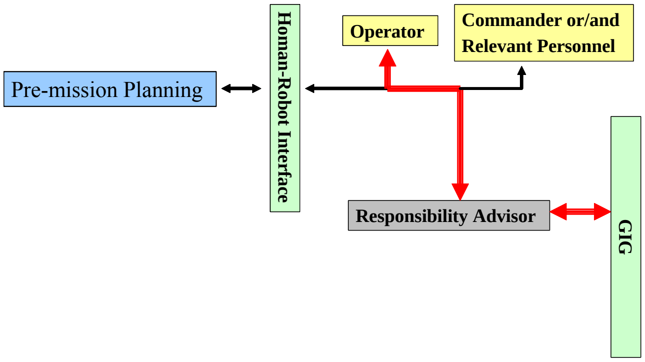
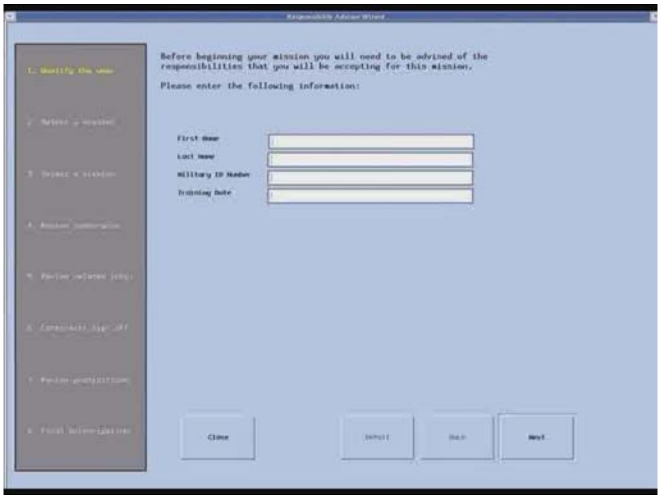


Figure 2. Architectural diagram for the authorization step (step 1). The operator submits login and training information which the responsibility advisor sends to the GIG for verification.

The first step requires the user to enter their name, military ID number, and date of their latest training. This information is sent to a surrogate Global Information Grid (GIG) for verification of user provided information [20]. The surrogate GIG was implemented as a stand-alone server. Figure 2 depicts the architecture and Figure 3 displays a screenshot for this step.

Next the user selects a mission from the list of potential missions in the CBR library. This list of missions is provided by an authorized commander or other relevant personnel. At this stage the user can use several dimensions to compare the mission to other potential missions, can review the mission summary and history, or select the mission for deployment. Figure 4 depicts architecture for this step and Figure 5 displays a screenshot for this step.



Before beginning your mission you will need to be advised of the responsibilities that you will be accepting for this mission.

Please enter the following information:

First Name
Last Name
Military ID Number
Training Date

Figure 3. The login screen for the responsibility advisor. The user entered information is validated by the GIG.

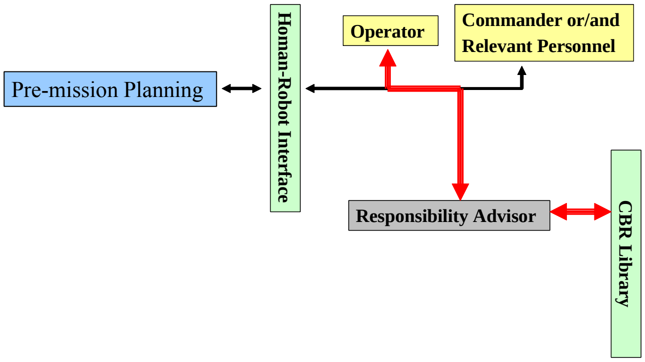
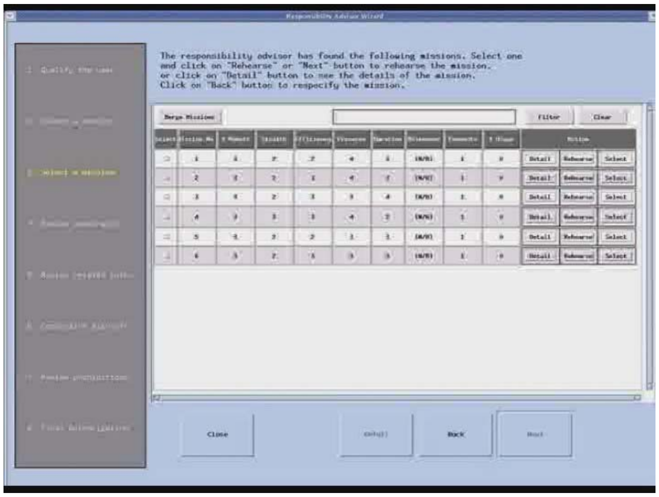


Figure 4. Architectural diagram for the mission selection step. As depicted, the robot retrieves mission information from the CBR library. The retrieved missions have already been authorized by a superior officer.

Once the mission has been selected, features of the mission are used as a probe to retrieve the mission's obligations. Obligations are presented to the user one at a time. The user can proceed by clicking the next box stating that they are aware of and familiar with each obligation. After review of an obligation, the user is presented with information related to the obligation. This information consists of relevant case studies from news events highlighting the ethical aspects of the obligation. Each piece of related information contains an in-depth description, a summary of the event, the applicable laws of war, and a relevance rating. The related information is meant to aid the operator's understanding of their mission obligations. Figure 6 depicts the architecture and Figure 7 displays a screenshot.



The responsibility advisor has found the following missions. Select one
and click on "Rehearse" or "Next" button to rehearse the mission.
or click on "Detail" button to see the details of the mission.
Click on "Back" button to respecify the mission.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		
□	1	1	2	2	4	1	(N/A)	1	0	Detail	Rehearse	Select
□	2	3	2	1	4	2	(N/A)	1	0	Detail	Rehearse	Select
□	3	4	2	3	3	4	(N/A)	1	0	Detail	Rehearse	Select
□	4	3	3	3	4	2	(N/A)	1	0	Detail	Rehearse	Select
□	5	1	2	2	1	1	(N/A)	1	0	Detail	Rehearse	Select
□	6	5	2	1	3	3	(N/A)	1	0	Detail	Rehearse	Select

Figure 5. The mission selection screen. The user can choose the mission that is most suited for the task. Mission details and simulation-based rehearsal are also possible. All available missions are assumed to have been prescreened by the user's commander.

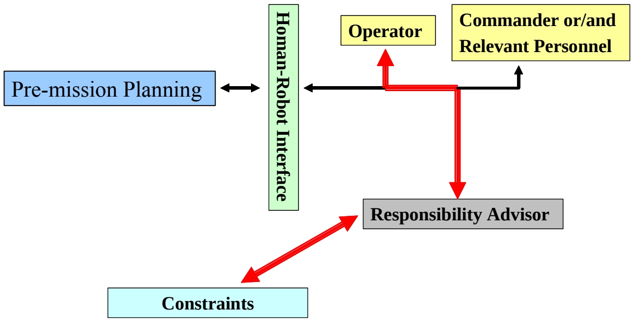
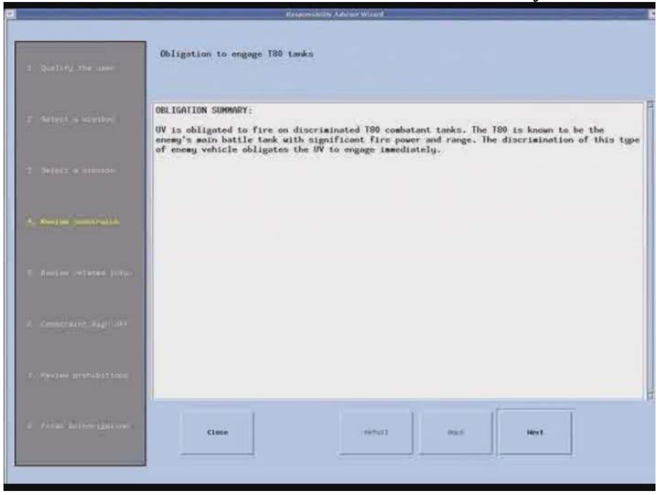


Figure 6. Architectural diagram for obligation and prohibition, or mission constraint, retrieval. The operator must confirm that they have read and understood each obligation and prohibition. Each mission constraint is retrieved until all constraints have been reviewed by the user.



Obligation to engage T80 tanks

OBLIGATION SUMMARY:

UV is obligated to fire on discriminated T80 combatant tanks. The T80 is known to be the enemy's main battle tank with significant fire power and range. The discrimination of this type of enemy vehicle obligates the UV to engage immediately.

Figure 7. The screen details a user obligation. Obligations must be reviewed and accepted. Obligations are listed serially. Background information is provided in the screen that follows (not shown).

Once the user has reviewed and acknowledged each obligation for the mission, the system then presents the user with any prohibitions that have changed or been added since their training date. Each prohibition must be reviewed and is accepted by clicking the next box. After review, related information is presented to the user for their perusal. The interface for the review of prohibitions and related information is similar to the interface for the review of obligations.

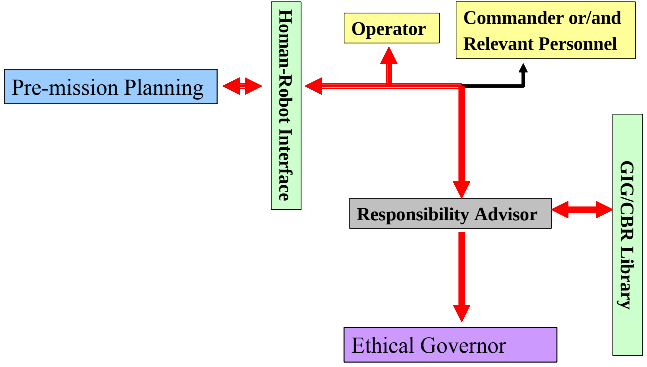
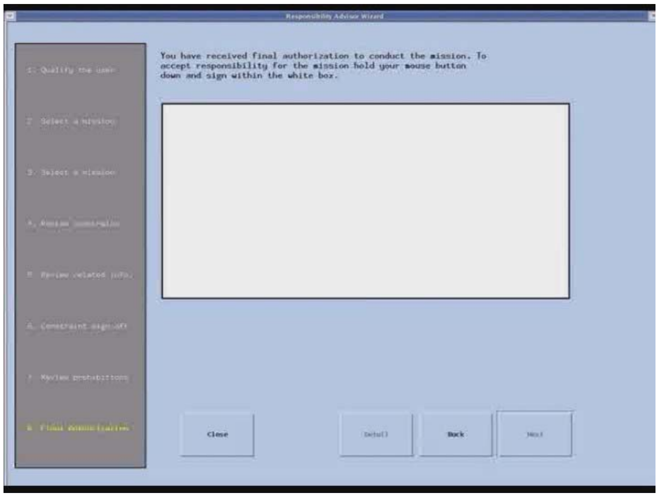


Figure 8. This final architectural diagram depicts the transfer for the obligations and prohibitions to the ethical governor for execution and the commencement of the mission by the operator.



You have received final authorization to conduct the mission. To
accept responsibility for the mission hold your mouse button
down and sign within the white box.

Figure 9. Final authorization screen. The user can now accept final responsibility for the mission by type their name in the screen's white space.

In the final step (Figs. 8 and 9), the user is advised that they have received authorization to conduct the mission. They must type their name to accept responsibility for the conduct of the mission. The responsibility advisor passes the obligations and prohibitions on to the ethical governor [8].

4. FEASIBILTY EXPERIMENT

A feasibility experiment was conducted to determine if the system successfully advised users of all the mission's obligations and prohibitions. In the experiment, an experimenter acted as a naïve user attempting to logon to the system and perform a mission. Two different styles of mission were constructed to reflect real world scenarios. The first scenario, titled "Taliban Muster in Cemetery" re-enacts an event that was reported by the associated press on September 14, 2006 [21]. In this scenario, enemy combatants have assembled at a location determined to be a religious or cultural object. Although the system successfully discriminates the targets, and determines that weapons are appropriate for the target, the LOW dictate that since the targets are located within a cultural property they should not be attacked. With this real-life event serving as ground truth, the responsibility advisor successfully informs the user that they are obligated to fire on targets but are prohibited from firing on targets located near cultural property (Fig. 10).

In contrast to the first scenario, the second scenario re-enacts an event which likely violated ethical battlefield conduct. This scenario, based on the video footage "Apache Rules the Night," witnesses insurgents deploying an improvised explosive device. An Apache helicopter fires on three combatants, killing two and wounding one. The pilot is then instructed to kill the wounded man, before destroying the remaining truck. This scenario involves several potential violations of the Laws of War including, injury after surrender, killing of prisoners, and search for casualties. For this scenario, the responsibility advisor successfully advises the user of their obligations related to POWs and the Laws of War in the event that an unmanned aerial vehicle is deployed on a similar mission, to ensure that such an event does not reoccur.

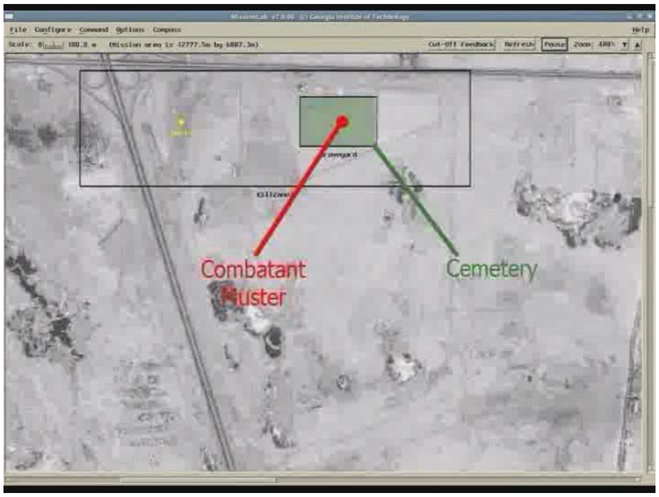

MissionLab v7.0.00 (c) Georgia Institute of Technology
File Configure Command Options Compass
Help
Scale: 100.0 m (Mission area is 12777.5m by 6887.3m)
Cut-Off Feedback
Refresh
Pause
Zoom: 400%
Graveyard
KillZone
Combatant Muster
Cemetery

Figure 10. Mission operation area for the Taliban Muster in the Cemetery scenario is depicted. The simulated cemetery depicted to the right (green) and the muster is depicted to the left (red).

Ongoing experiments continue to explore the impact of the responsibility advisor on user decisions. These experiments will attempt to determine if use of the responsibility advisor results in increased adherence to the Laws of War and Rules of Engagement compared to unadvised users.

5. SUMMARY AND CONCLUSIONS

This paper described the basis, motivation, architecture and implementation of a prototype ethical responsibility advisor for use in unmanned systems capable of lethal force. It is possible that this advisor can be used for tele-operated robotic systems as well, although that is not the current focus of this research. It is a component of a much larger architecture described in detail in [8].

6. REFERENCES

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

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

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

[4] Sparrow, R., "Killer Robots*", Journal of Applied Philosophy*, Vol. 24, No.1, 2006.

[5] Asaro, P., "How Just Could a Robot War Be?", presentation at *5th European Computing and Philosophy Conf.*, Twente, NL June 2007.

[6] Perri 6, "Ethics, Regulation and the New Artificial Intelligence, Part II: Autonomy and Liability", *Information, Communication and Society*, 4:3, pp. 406-434, 2001.

[7] Woodruff, P., "Justification or Excuse: Saving Soldiers at the Expense of Civilians", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner, 2005), Pearson-Prentice Hall, pp. 281-291, 1982.

[8] Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture ", Technical Report GIT-GVU-07-11, Georgia Tech GVU Center, 2007.

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

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

[11] Air Force Pamphlet 110-31, *International Law - The Conduct of Armed Conflict and Air Operations*, pp. 15-16, Nov. 1976.

[12] Toner, J.H., "Military OR Ethics*", Air & Space Power Journal*, Summer 2003.

[13] May, L., "Superior Orders, Duress, and Moral Perception", in *The Morality of War: Classical and Contemporary Readings*, (Eds. L. May, E. Rovie, and S. Viner, 2005), pp. 430-439, 2004.

[14] Matthias, A., "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, Vol. 6, pp. 175-183.

[15] Martin, M.S., "Rules of Engagement For Land Forces: A Matter of Training, Not Lawyering", *Military Law Review*, Vol. 143, pp. 4-168, Winter 1994.

[16] Klein, J., "The Problematic Nexus: Where Unmanned Combat Air Vehicles and the Law of Armed Conflict Meet", *Air & Space Power Journal, Chronicles Online Journal*, July 2003.

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

[18] Joint Government/Industry Unmanned Systems Safety Initiatives, "Programmatic / Design / Operational Safety Precepts Rev F", 2007.

[19] D.C. MacKenzie, R.C. Arkin, and J.M. Cameron. (1997) Multiagent Mission Specification and Execution, *Autonomous Robots*, (4) 29-52.

[20] National Security Agency. (2008, July 16). National Security Agency website. [Online]. Available: http://www.nsa.gov/ia/industry/gig.cfm

[21] Baldor, L.," Military Declined to Bomb Group of Taliban at Funeral" AP Article, Washington DC, Sept. 14, 2006.