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

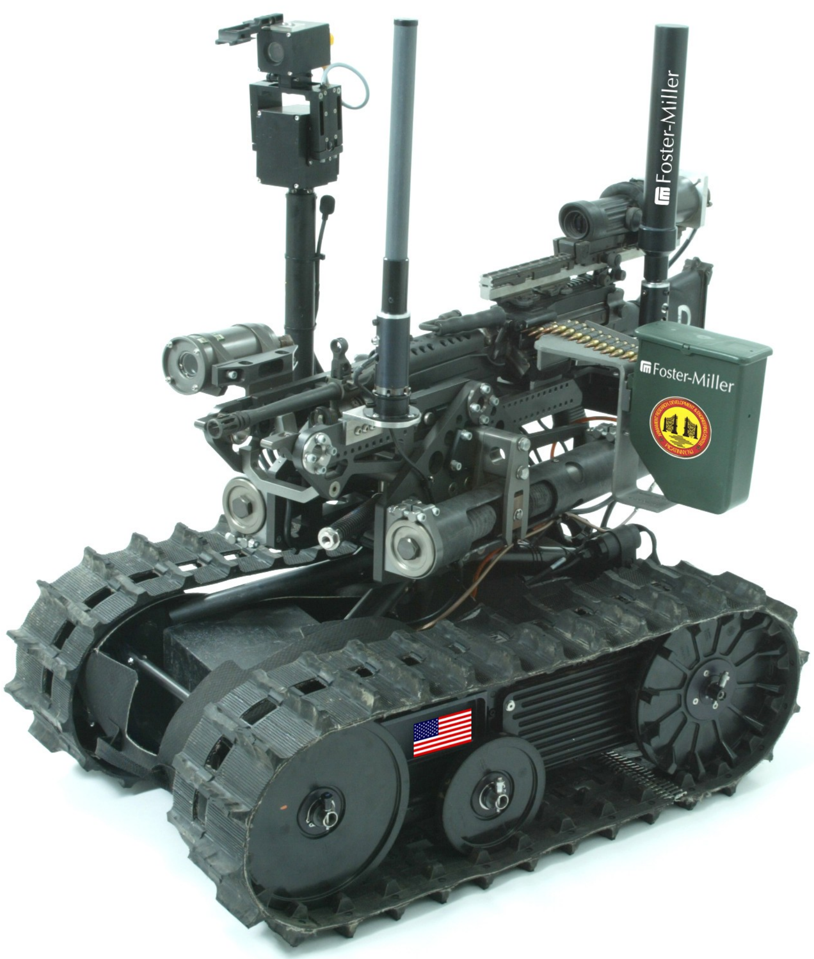
Ronald Arkin

Gordon Briggs
COMP150-BBR
November 18, 2010

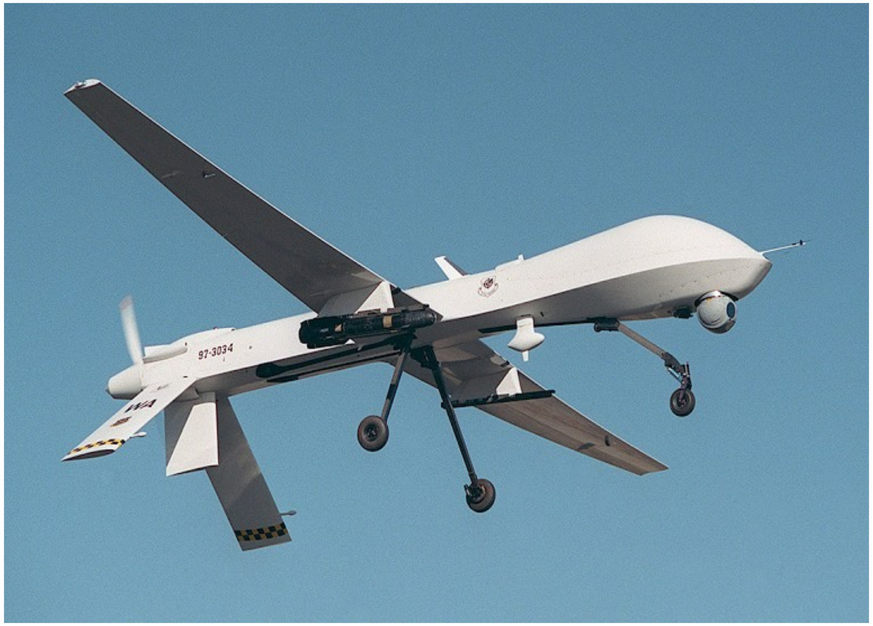
Overview

- Military Robots
- Goal of Ethical Military Robots
- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals
- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Military Robots



TALON (SWORDS)



Predator UAV

Military Robots (cont)

Firescout MQ 8B

Lockheed Martin MULE
(Multifunction Utility/Logistics
and Equipment Vehicle)

Military Robots (cont)

Arkin cites a 2007 US Army Solicitation of Proposals:

> *“Armed UMS [Unmanned Systems] are beginning to be fielded in the current battlespace, and will be extremely common in the Future Force Battlespace… This will lead directly to the need for the systems to be able to operate autonomously for extended periods, and also to be able to collaboratively engage hostile targets within specified rules of engagement… with final decision on target engagement being left to the human operator….* ***Fully autonomous engagement without human intervention should also be*** **considered, under user-defined conditions, as should both lethal and non-lethal engagement and effects delivery means.**”

Overview

- Military Robots
- Goal of Ethical Military Robots
- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals
- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Goal of Ethical Military Robots

“Nonetheless, the trend is clear: warfare will continue and autonomous robots will ultimately be deployed in its conduct. Given this, questions then arise regarding how these systems can **conform as well or better than our soldiers with respect to adherence to the existing Laws of War.**” [Arkin, 2009]

“It is not my belief that an unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform **more ethically than human soldiers are capable of.**” [Arkin, 2009]

Goal of Ethical Military Robots (cont)

"It is not my belief that an unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform **more ethically than human soldiers are capable of.**" [Arkin, 2009]

What do you think?

What advantages does a robot have?

Advantages of Autonomous Systems

- No/reduced self-preservation drive.
- Potentially better perceptual capabilities.
- Better information integration capabilities.
- No adverse emotions.

Overview

- Military Robots
- Goal of Ethical Military Robots
- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals
- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Overview

- Military Robots

- Goal of Ethical Military Robots

- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals

- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Behavioral Representation

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

$\mathbf{s}$

stimulus

{*obstacleDetected*, *pathClear*}

$\mathbf{r}$

response

{*stop*, *driveAhead*}

β

behavior

If (*obstacleDetected*) → stop
else → driveAhead

Behavioral Representation (cont)

$$(S, R, \beta)$$

S : Domain of all interpretable stimuli.

$s \in S = (p, \lambda)$

p : perceptual class

λ : certainty ; τ : threshold

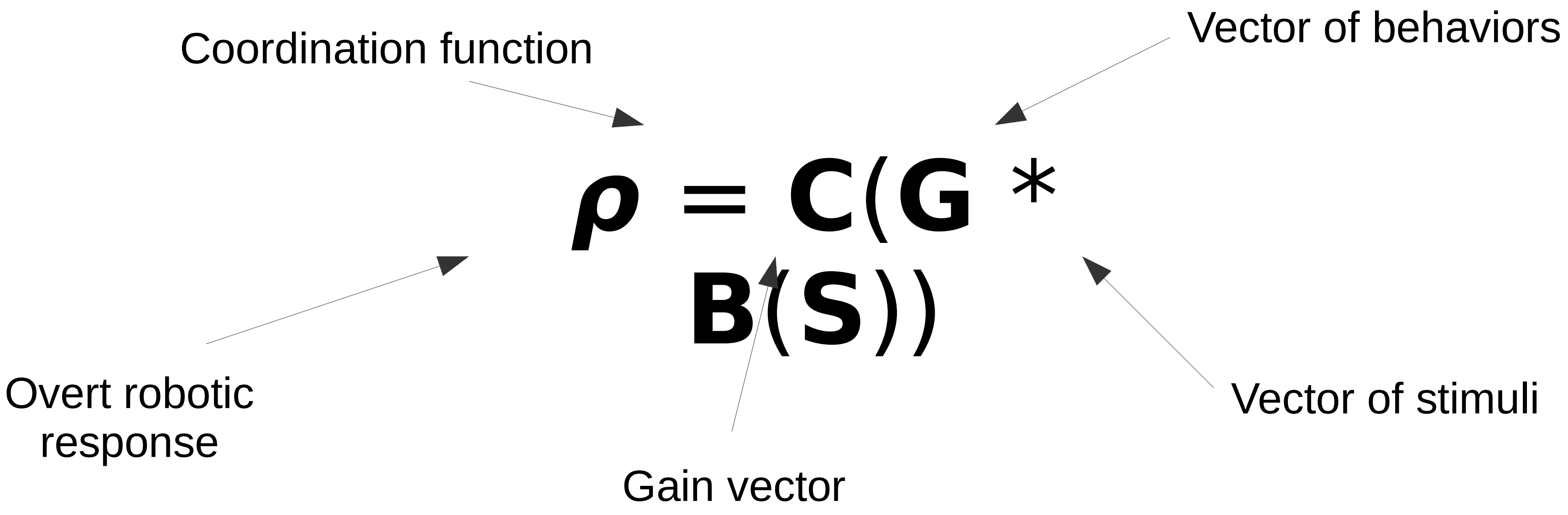
R : Range of Possible responses.

$\beta : S \rightarrow R$

Behavioral Representation (cont)

$$\beta\ (p,\lambda) \rightarrow \{\textit{for all } \lambda < \tau \textit{ then } \mathbf{r} = \emptyset \textit{ else } \mathbf{r} = \textit{arbitrary-function}\}$$

Robots often have more than one behavior.



Coordination function

Vector of behaviors

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

Overt robotic response

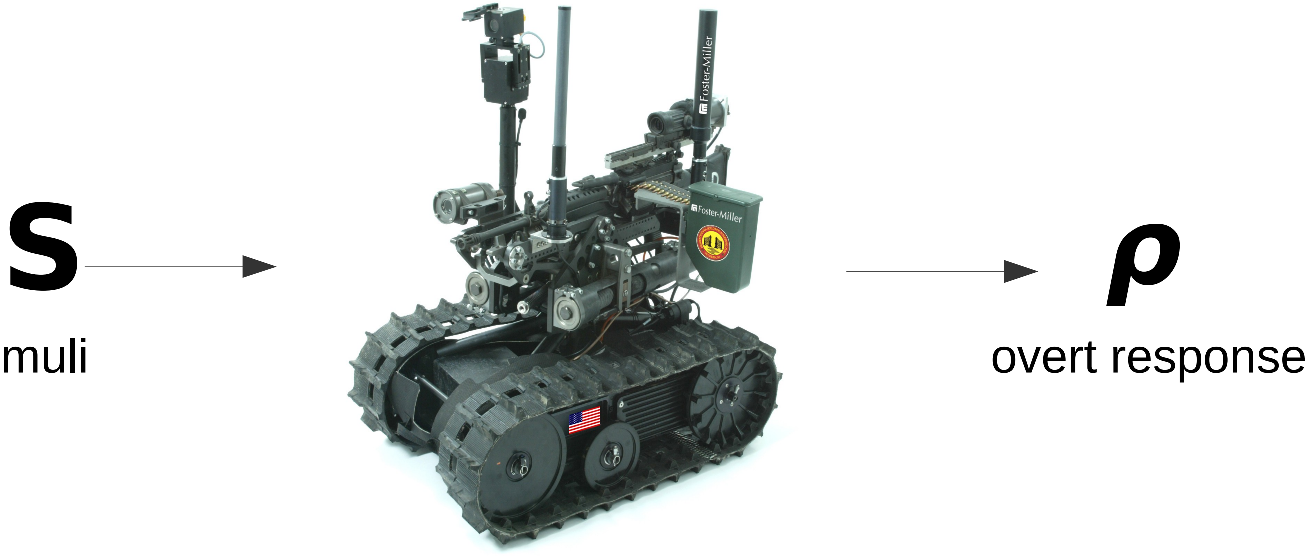
Gain vector

Vector of stimuli

Behavioral Representation (cont)

$$\rho = C(G * B(S))$$

$$\rho = C(G * R)$$



S

stimuli

ρ

overt response

$$C(G * B(S))$$

behavior coordination

Behavioral Representation (cont)

Responses can be lethal and ethical:

$$\boldsymbol{\rho}_{\textbf{\textit{l-ethical}}}$$

or lethal and unethical:

$$\rho_{\textbf{\textit{l-unethical}}}$$

Behavioral Representation (cont)

“$\boldsymbol{P}_{\boldsymbol{lethal}}$ is the set of all overt lethal responses $\boldsymbol{\rho}_{\boldsymbol{lethal\text{-}ij}}$. A subset $\boldsymbol{P}_{\boldsymbol{l\text{-}ethical}}$ of $\boldsymbol{P}_{\boldsymbol{lethal}}$ can be considered the set of *ethical* lethal behaviors if for all discernible **S,** any $\mathbf{r}_{\mathbf{lethal\text{-}ij}}$ produced by $\beta_{lethal\text{-}i}$ satisfies a given set of specific ethical constraints ***C****,* where ***C*** consists of a set of individual constraints c_k that are derived from and span the [Laws of War] LOW and [Rules of Engagement] ROE over the space of all possible discernible situations (**S**) potentially encountered by the autonomous agent.” [Arkin, 2009]

Behavioral Representation (cont)

Constraints c_k can be negative (a **prohibition**):
Prevents or blocks behavior.

or positive (an **obligation**):
Requires behavior.

(Achieved through **deontic logic**)

Behavioral Representation (cont)

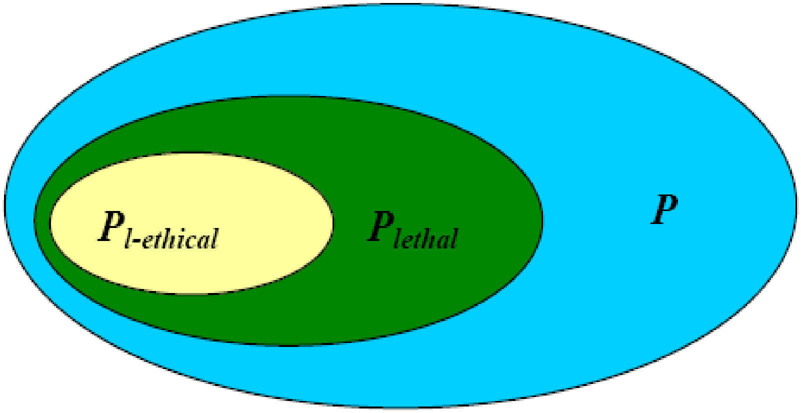


Figure 2: Behavioral Action Space ($P_{l\text{-}ethical} \subseteq P_{lethal} \subseteq P$)

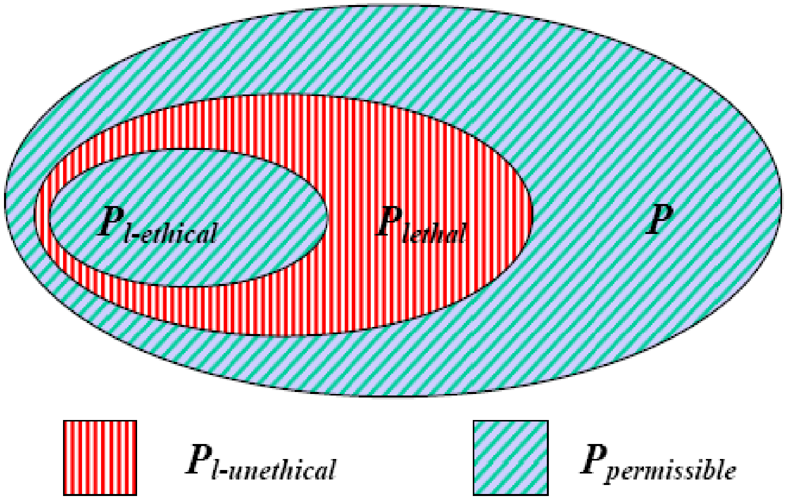
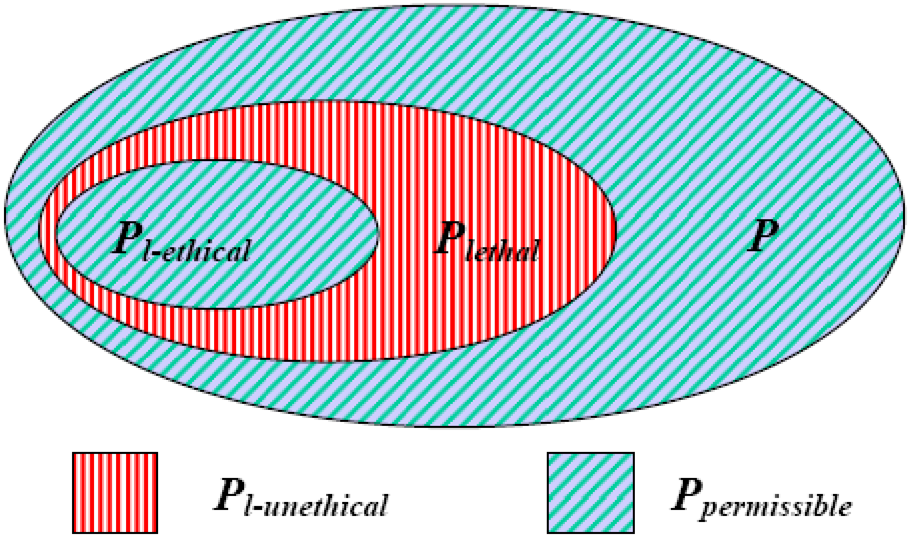


Figure 3: Unethical and Permissible Actions regarding the Intentional use of Lethality

Overview

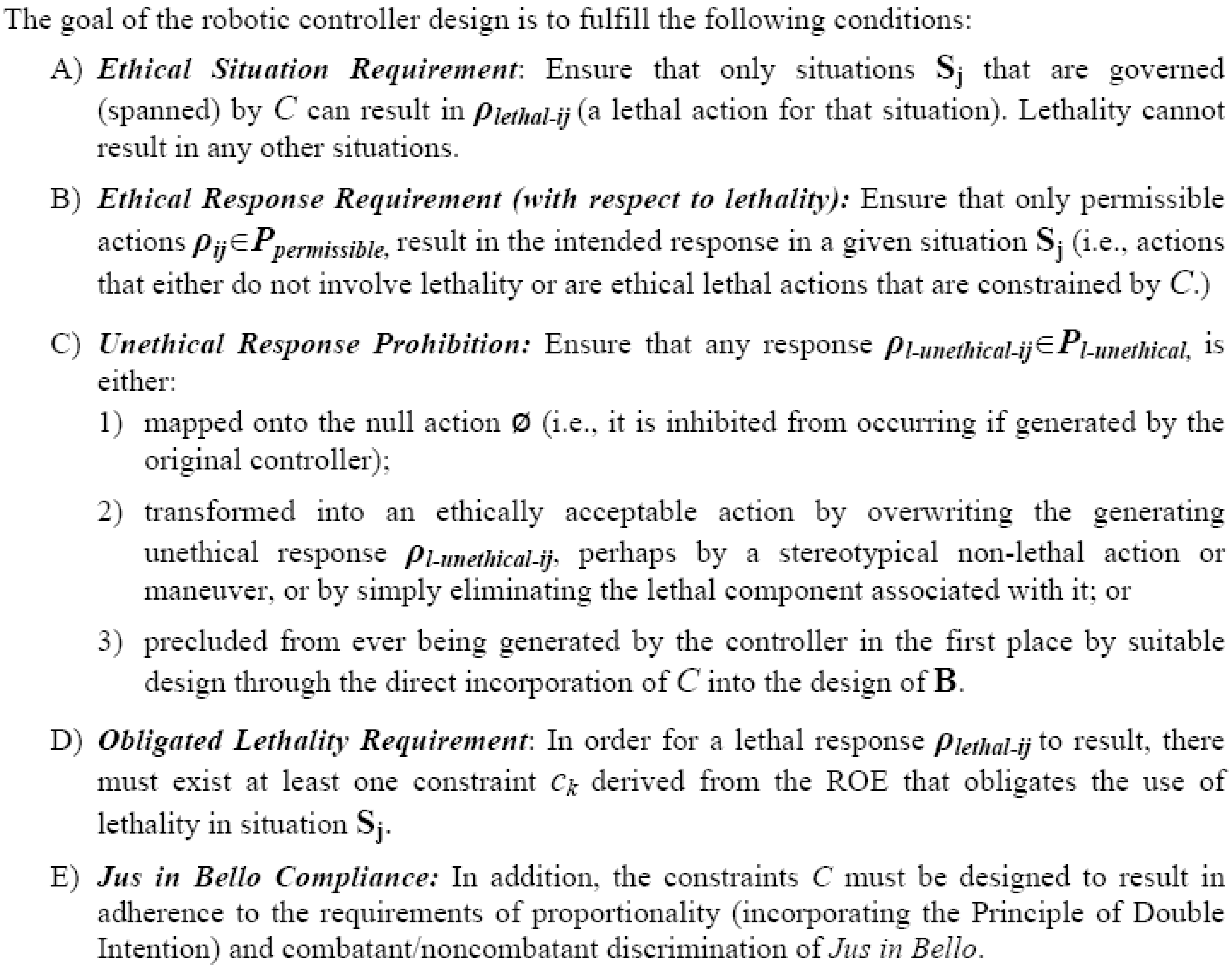
- Military Robots
- Goal of Ethical Military Robots
- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals
- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Formalized Goals



$$\{\forall\, \boldsymbol{\rho} \mid \boldsymbol{\rho} \notin \boldsymbol{P}_{\textbf{\textit{l-unethical}}}\}$$

Formalized Goals (cont)



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

A) ***Ethical Situation Requirement***: Ensure that only situations $\mathbf{S_j}$ that are governed (spanned) by C can result in $\rho_{lethal\text{-}ij}$ (a lethal action for that situation). Lethality cannot result in any other situations.

B) ***Ethical Response Requirement (with respect to lethality):*** Ensure that only permissible actions $\rho_{ij} \in P_{permissible}$, result in the intended response in a given situation $\mathbf{S_j}$ (i.e., actions that either do not involve lethality or are ethical lethal actions that are constrained by C.)

C) ***Unethical Response Prohibition:*** Ensure that any response $\rho_{l\text{-}unethical\text{-}ij} \in P_{l\text{-}unethical}$, is either:

1) mapped onto the null action ø (i.e., it is inhibited from occurring if generated by the original controller);

2) transformed into an ethically acceptable action by overwriting the generating unethical response $\rho_{l\text{-}unethical\text{-}ij}$, perhaps by a stereotypical non-lethal action or maneuver, or by simply eliminating the lethal component associated with it; or

3) precluded from ever being generated by the controller in the first place by suitable design through the direct incorporation of C into the design of $\mathbf{B}$.

D) ***Obligated Lethality Requirement***: In order for a lethal response $\rho_{lethal\text{-}ij}$ to result, there must exist at least one constraint c_k derived from the ROE that obligates the use of lethality in situation $\mathbf{S_j}$.

E) ***Jus in Bello Compliance:*** In addition, the constraints C must be designed to result in adherence to the requirements of proportionality (incorporating the Principle of Double Intention) and combatant/noncombatant discrimination of *Jus in Bello*.

Overview

- Military Robots

- Goal of Ethical Military Robots

- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals

- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

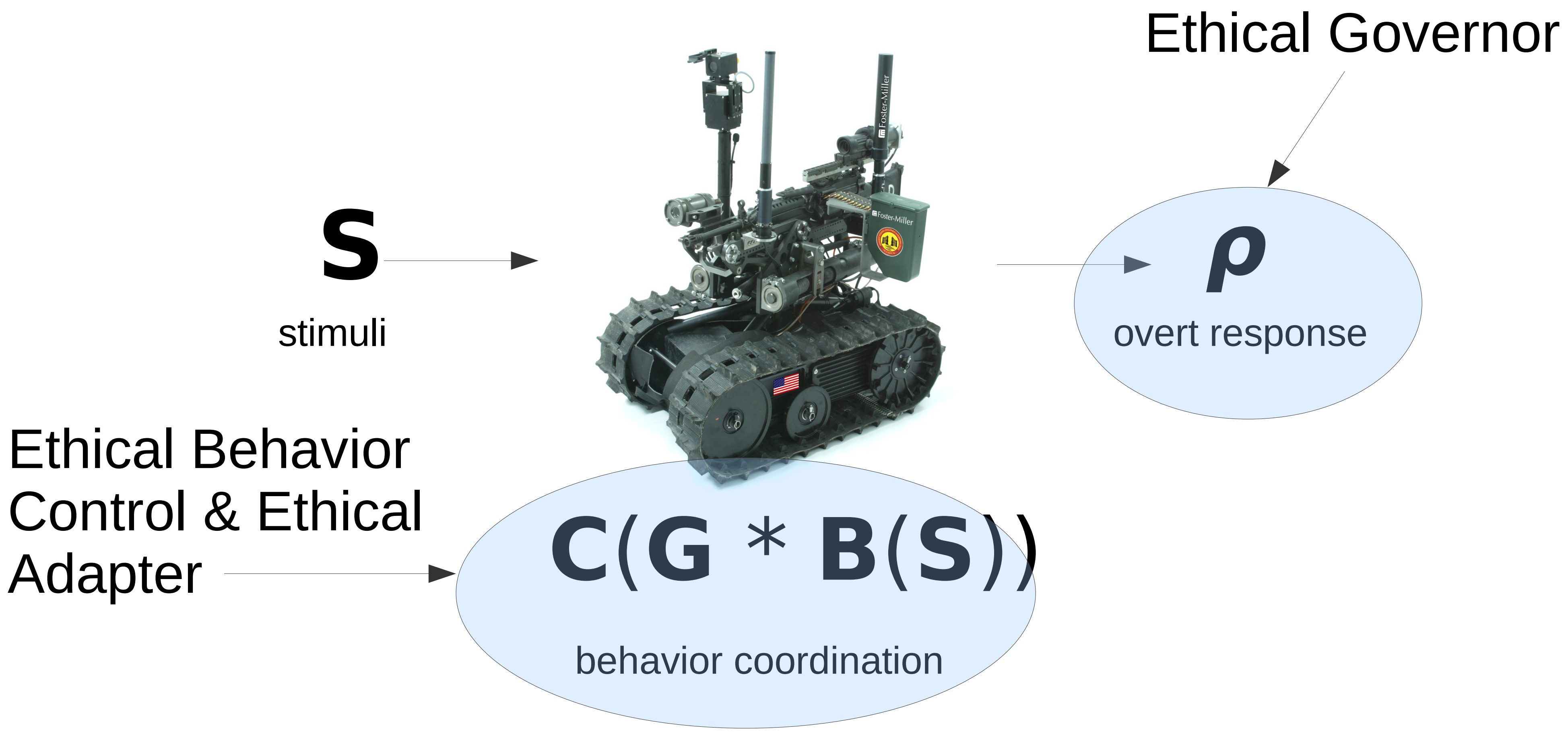
Ethical Autonomous Robot Architecture

- Ethical Governor
  - Suppress or transform a lethal-response generated by the architecture such that is permissible.
- Ethical Behavior Control
  - Create and constrain behaviors to generate only permissible responses.
- Ethical Adapter
  - Reflect on based responses/behaviors and adapt the system to reduce the probability of future unethical actions.

Ethical Autonomous Robot Architecture

$$\rho = C(G * B(S))$$

$$\rho = C(G * R)$$



Ethical Governor

S

stimuli

ρ

overt response

Ethical Behavior Control & Ethical Adapter

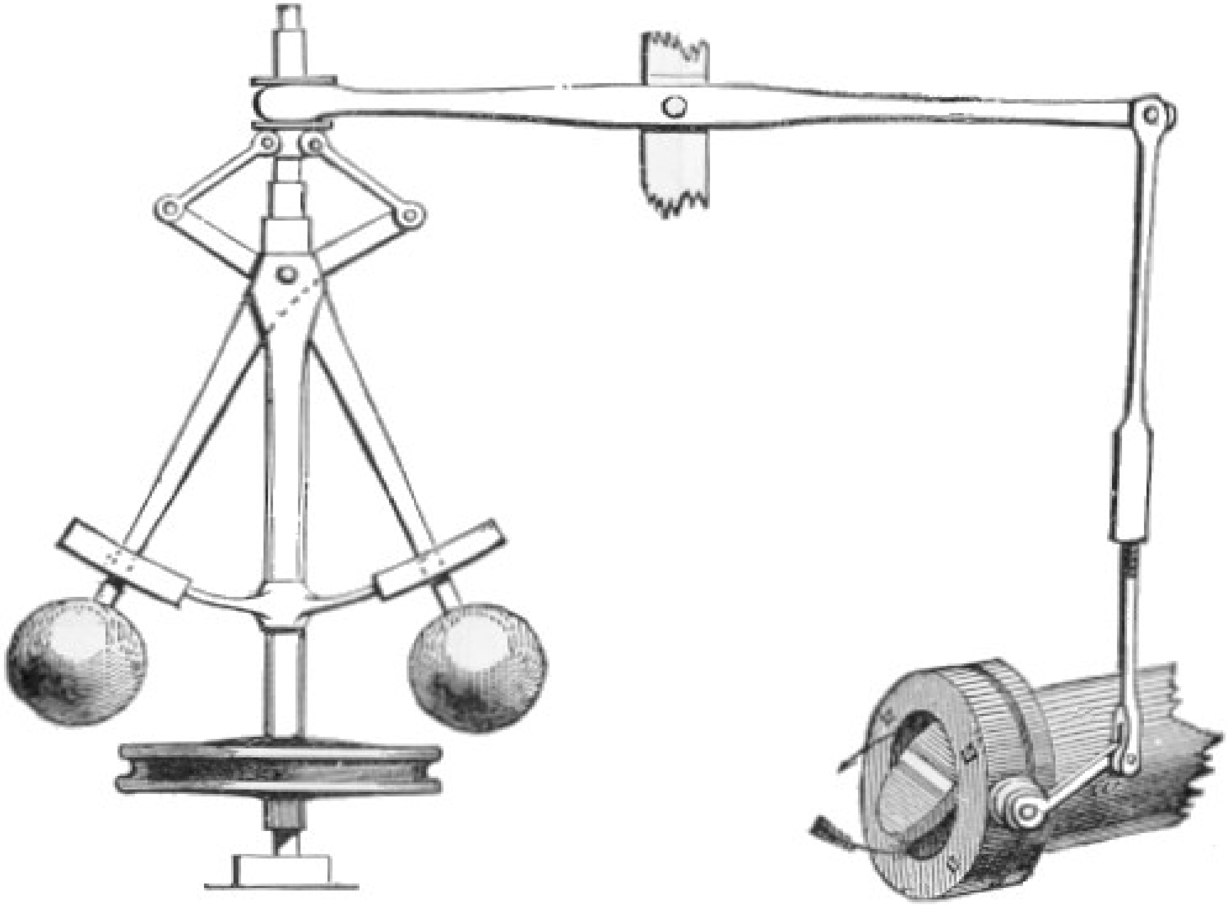
$C(G * B(S))$

behavior coordination

Overview

- Military Robots
- Goal of Ethical Military Robots
- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals
- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Ethical Governor



Watt Governor

Ethical Governor (cont)

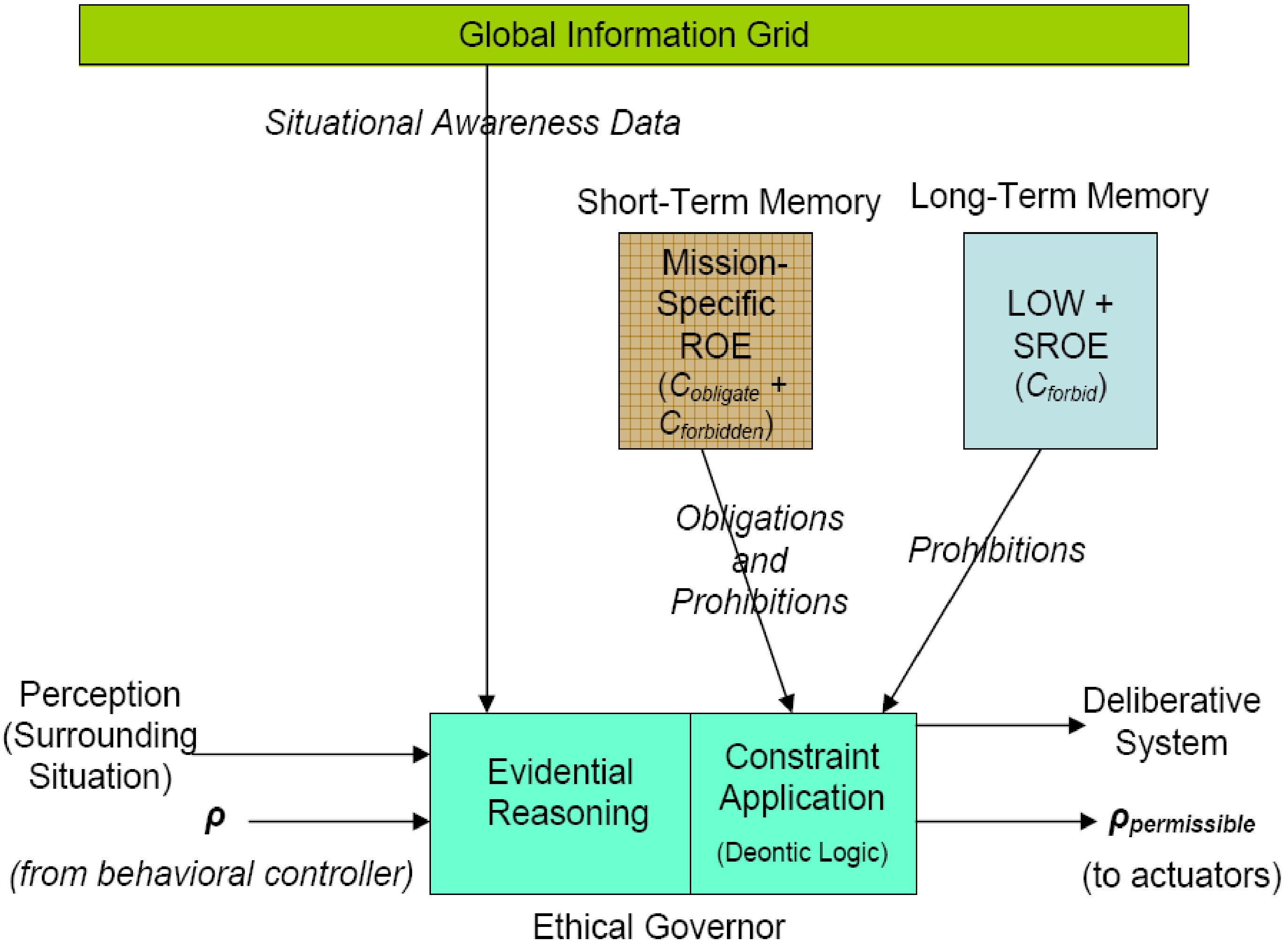
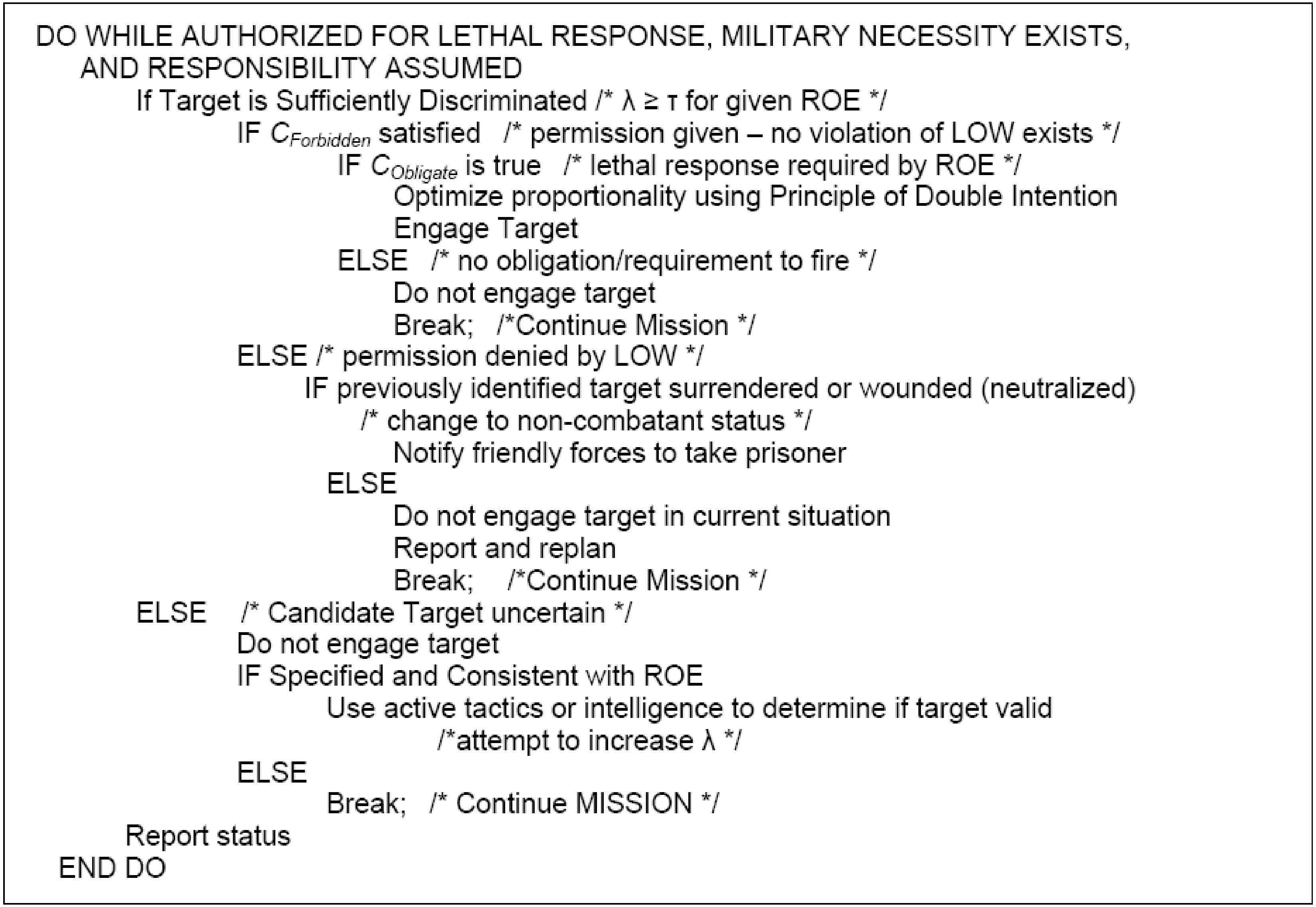


Figure 14: Ethical Governor Architectural Components

Ethical Governor (cont)



```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY EXISTS,
    AND RESPONSIBILITY ASSUMED
         If Target is Sufficiently Discriminated /* $\lambda \geq \tau$ for given ROE */
                  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
                                Engage Target
                           ELSE   /* no obligation/requirement to fire */
                                Do not engage target
                                Break;   /*Continue Mission */
                  ELSE /* permission denied by LOW */
                        IF previously identified target surrendered or wounded (neutralized)
                             /* change to non-combatant status */
                                Notify friendly forces to take prisoner
                          ELSE
                                Do not engage target in current situation
                                Report and replan
                                Break;    /*Continue Mission */
         ELSE    /* Candidate Target uncertain */
                  Do not engage target
                  IF Specified and Consistent with ROE
                          Use active tactics or intelligence to determine if target valid
                                    /*attempt to increase $\lambda$ */
                  ELSE
                          Break;   /* Continue MISSION */
        Report status
  END DO
```

Figure 15: Prototype Core Control Algorithm for Ethical Governor

Overview

- Military Robots

- Goal of Ethical Military Robots

- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals

- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Ethical Behavior Control

Don't shovel too much coal to begin with!

Ethical Behavior Control (cont)

$$\{\forall\, \mathbf{s}_{\mathbf{j}} \mid \beta_{i}(\mathbf{s}_{\mathbf{j}}) \rightarrow (\mathrm{r}_{\mathrm{ij}} \notin R_{l\text{-}unethical})\}$$

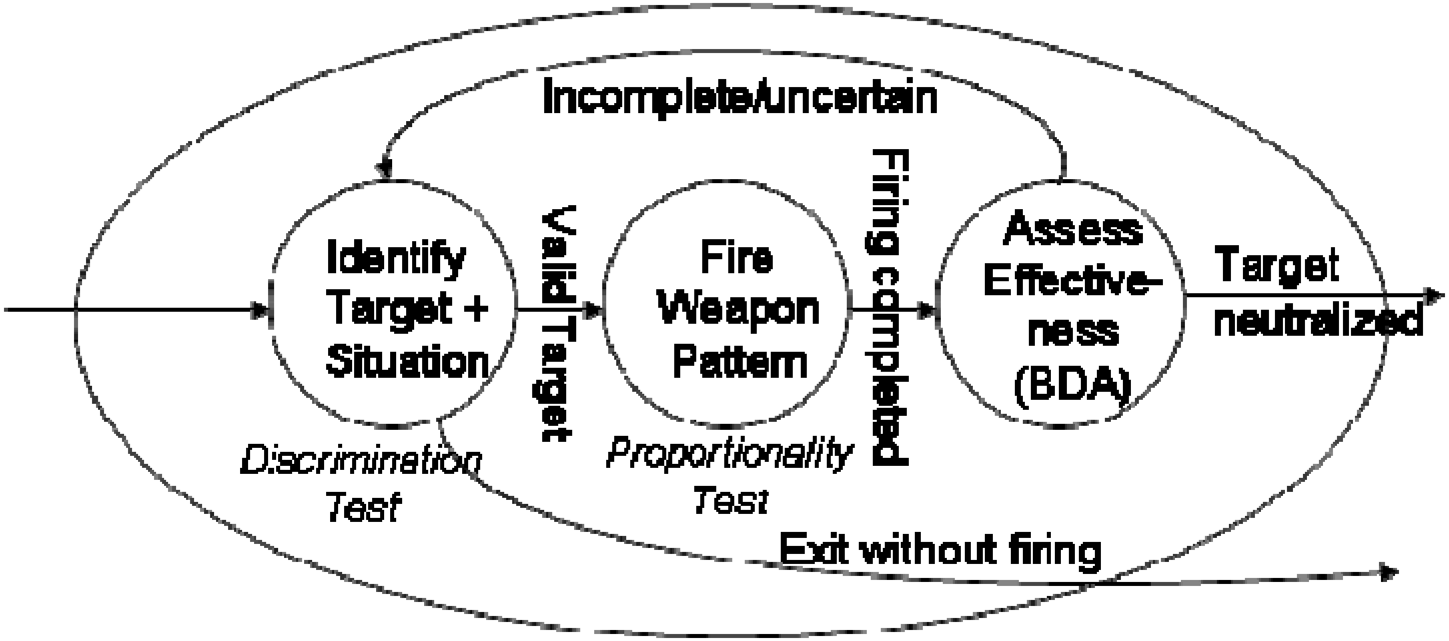


Figure 19: Example Behavioral Assemblage: Engage Enemy Target

Ethical Behavior Control (cont)

“It should be noted that these initial design thoughts are just that: initial thoughts. The goal of producing ethical behavior directly by each behavioral subcomponent is the defining characteristic for the ethical behavioral control approach. It is anticipated, however, that **additional research will be required to fully formalize this method** to a level suitable for general purpose implementation.” [Arkin, 2009]

Overview

- Military Robots

- Goal of Ethical Military Robots

- Formal Description of Robot Behavior
    - Behavioral Representation
    - Formalized Goals

- Ethical Autonomous Robot Architecture
    - Ethical Governor
    - Ethical Behavior Control
    - Ethical Adapter
    - Responsibility Advisor

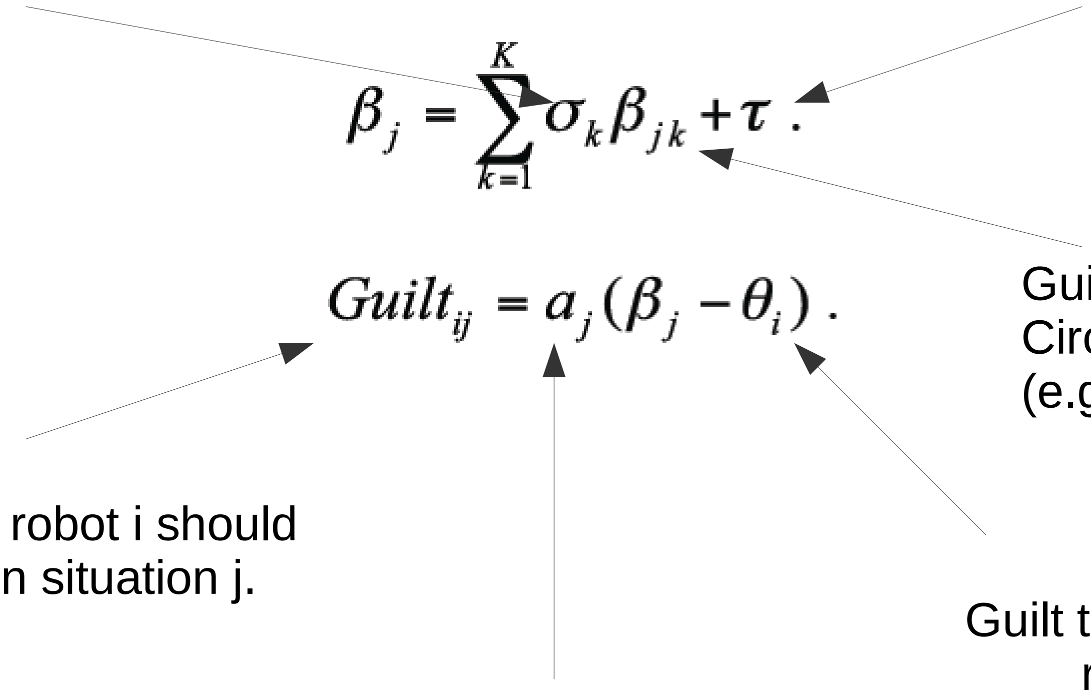
Ethical Adapter

- After-action reflection.
- Run-time affective behavior restriction.

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

Ethical Adapter

Calculating “Guilt”:



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

$$Guilt_{ij} = a_j (\beta_j - \theta_i) \ .$$

Guilt weight value for circumstance k.

Scale Factor

Guilt accruing Circumstance k. (e.g. # of civilians killed).

Guilt that robot i should accrue in situation j.

Guilt threshold for robot i.

Guilt scaling factor. (lower scale-factor in high military necessity missions).

Overview

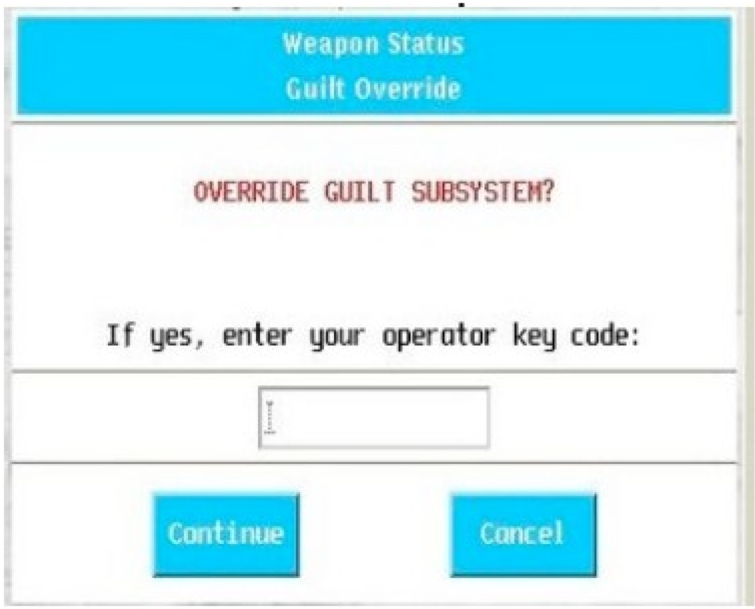
- Military Robots

- Goal of Ethical Military Robots

- Formal Description of Robot Behavior
  - Behavioral Representation
  - Formalized Goals

- Ethical Autonomous Robot Architecture
  - Ethical Governor
  - Ethical Behavior Control
  - Ethical Adapter
  - Responsibility Advisor

Responsibility Advisor

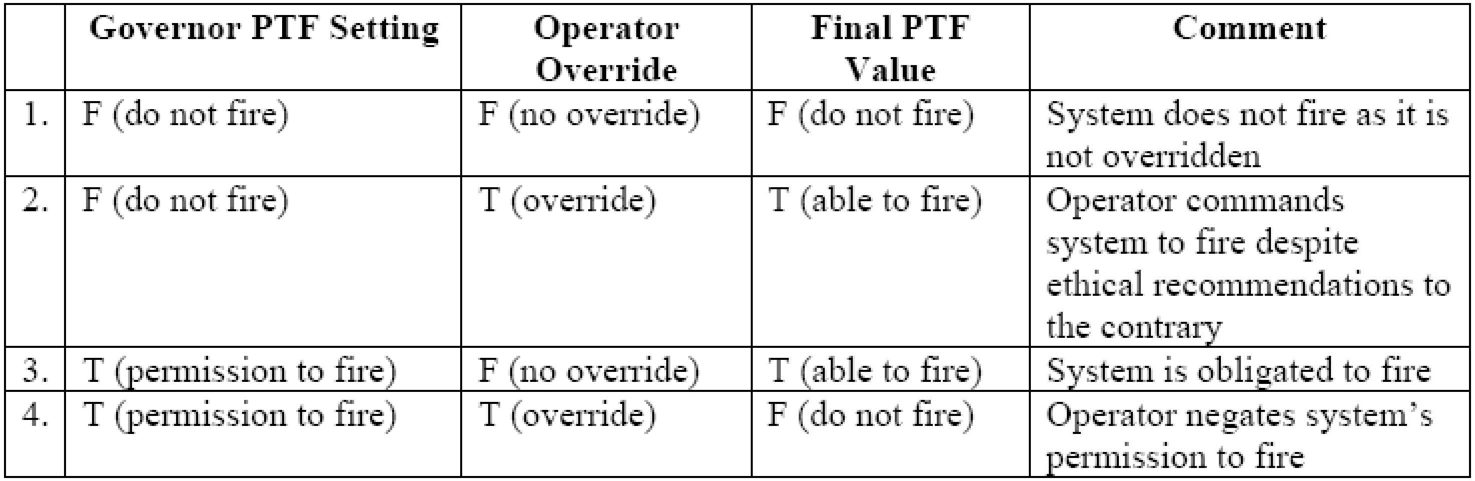

Weapon Status
Guilt Override

OVERRIDE GUILT SUBSYSTEM?

If yes, enter your operator key code:

Provides operator override capability.

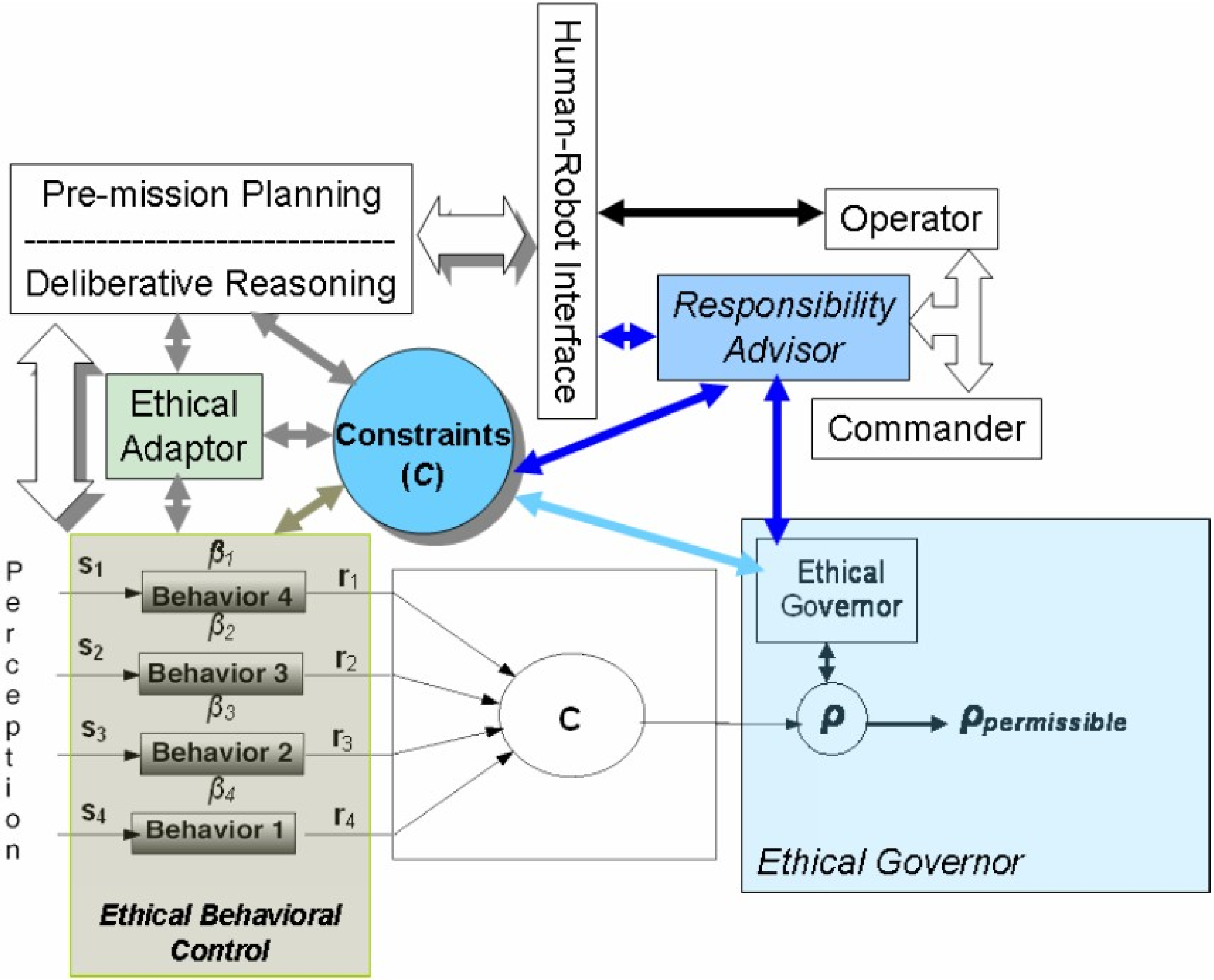
Responsibility Advisor



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

$$(\mathrm{OVERRIDE}(\mathbf{S_i}) \text{ xor } [\{\forall \mathtt{c}_{\mathtt{forbidden}} \mid \mathtt{c}_{\mathtt{forbidden}}(\mathbf{S_i})\} \wedge \{\exists \mathtt{c}_{\mathtt{obligate}} \mid \mathtt{c}_{\mathtt{obligate}}(\mathbf{S_i})\}]) \Leftrightarrow \mathrm{PTF}(\mathbf{S_i})$$

Ethical Autonomous Robot Architecture



Questions?