Australian Government
Department of Defence
Defence Science and
Technology Organisation

C^2 Design for Ethical Agency over Killing in War

Patrick Hew
POC: Patrick.Hew@dsto.defence.gov.au

15th International Command & Control Research & Technology Symposium
22-24 June 2010 (Santa Monica, California)

DSTO

Australian Government
Department of Defence
Defence Science and
Technology Organisation

Synopsis

- Develop a C^2 design for assigning ***ethical agency*** over killing in war
  - Integrate engineering *vs* philosophical notions of “autonomy”
- Establish that for foreseeable technology
  - **Necessary** for a human to be “on” the firing loop
  - **Neither necessary nor sufficient** for a human to be “in” the firing loop
- Reinvigorate robotics & automation for Western military forces via C^2 design

Australian Government
Department of Defence
Defence Science and
Technology Organisation

Outline

- A Question of Killer Robots
- Engineering vs Philosophical “Autonomy”
  - Intelligent Agents
  - Supervisory Control
- *Proposition*: The Ethical Agent
  - Rationale from Just War Theory
- Implications for C^2 Theory and Practice

Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Question of Killer Robots

- Western ethics on warfare require that someone be held responsible for the deaths that occur [Sparrow]
- ***Current axiom***: “Someone” is human being
- Systems engineering questions:
  - What properties of a human enable them to be held responsible?
  - Allocate activities to humans and/or machines?
  - If duties must be held by a human, what must be done to support the human in this capacity?

R Sparrow, “Killer Robots,” *Journal of Applied Philosophy*, vol. 24, no. 1, pp.62-77, 2007.

Relevance to Evolution of C^2

- Technology development for automated target recognition, “brilliant munitions” …
  - 1980s-90s Substantial efforts on expectation of high potential benefit
  - circa 2000 Research slowed on concerns of ethical accountability
  - Current Renewed interest to fix manpower footprint from unmanned systems
- Clarify debate on ethics of “killer robots”
  - Growth post-2001 in unmanned systems
  - “Killer robot” = “Brilliant munition” (or not)?

Australian Government
Department of Defence
Defence Science and
Technology Organisation

Intelligent Agents

- AI definition of *Intelligent Agent*

  "Autonomous entity that observes and acts upon an environment and directs its activity towards achieving goals."

- This is ***engineering autonomy***
  - Closing a loop from sensors to effectors

- No restrictions on an agents' construction
  - Humans, machines, organisations, …

S J Russell, P Norvig, *Artificial Intelligence: A Modern Approach*, 2nd Edition ed., Upper Saddle River, NJ: Prentice Hall, 2003.

Australian Government
Department of Defence
Defence Science and Technology Organisation

Supervisory Control

- Sheridan Model of *Supervisory Control*

  "One or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors to the controlled process or task environment."

  =

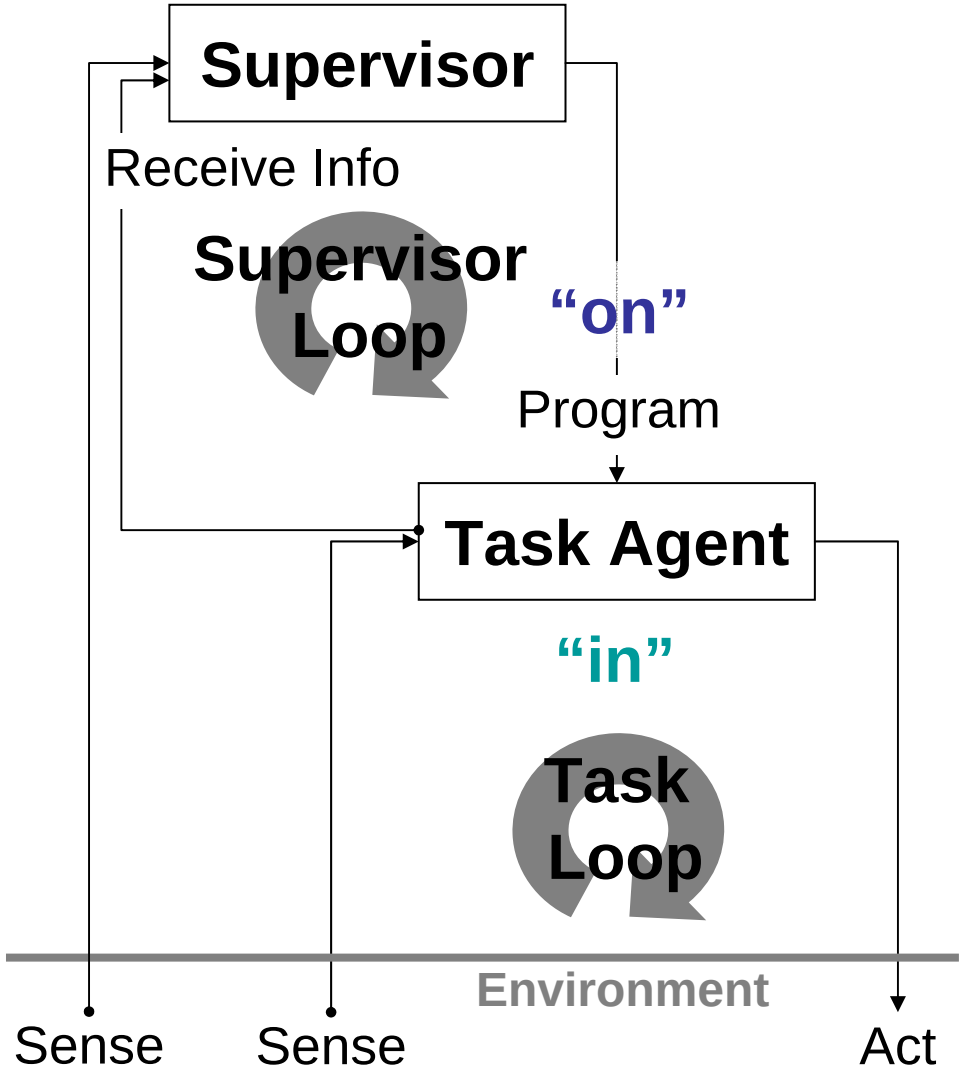
  "One or more operators are intermittently programming and receiving information from an artificial intelligent agent."

- *Informally*: "on" the loop
  - Versus human being "in" the control loop

T B Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, Cambridge: MIT Press, 1992.

Task and Supervisor Agents

- Task Agent
  - Sense & Act into environment
- Supervisor Agent
  - Sense from environment
  - Receive Info from Task Agent
  - Program Task Agent



Australian Government
Department of Defence
Defence Science and
Technology Organisation

Lethal Agents

- Lethal Agent
  - Particular form of Task Agent
  - Closes a firing loop from sensors to weapons

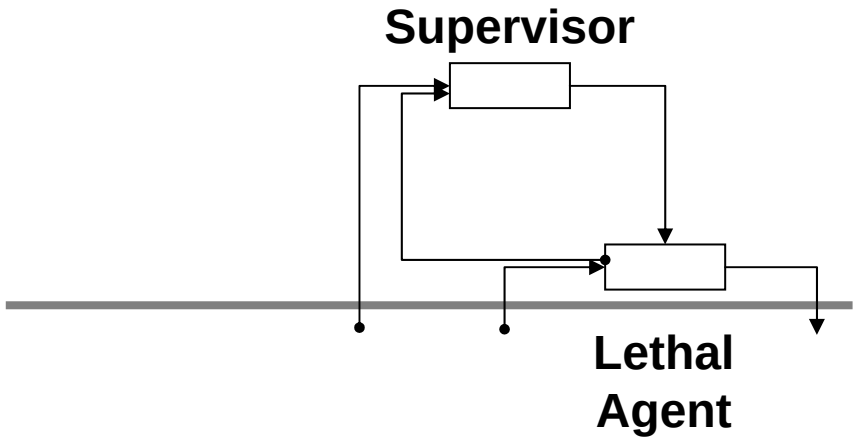
Australian Government
Department of Defence
Defence Science and
Technology Organisation

Engineering and Philosophy

- Lethal agents can be built from machines

so

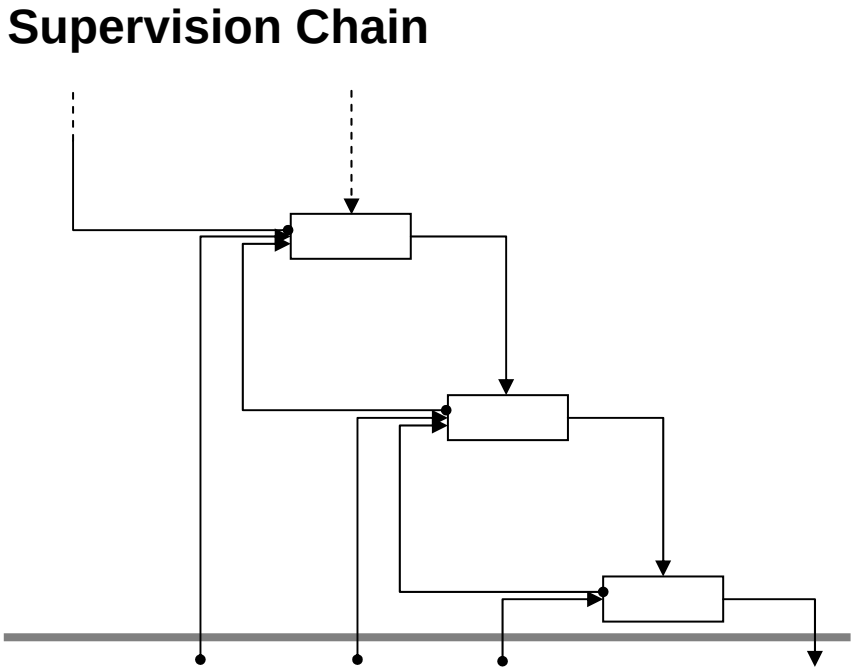
- Unique qualities of humans *vs* machines are in the structures for supervision



Australian Government
Department of Defence
Defence Science and
Technology Organisation

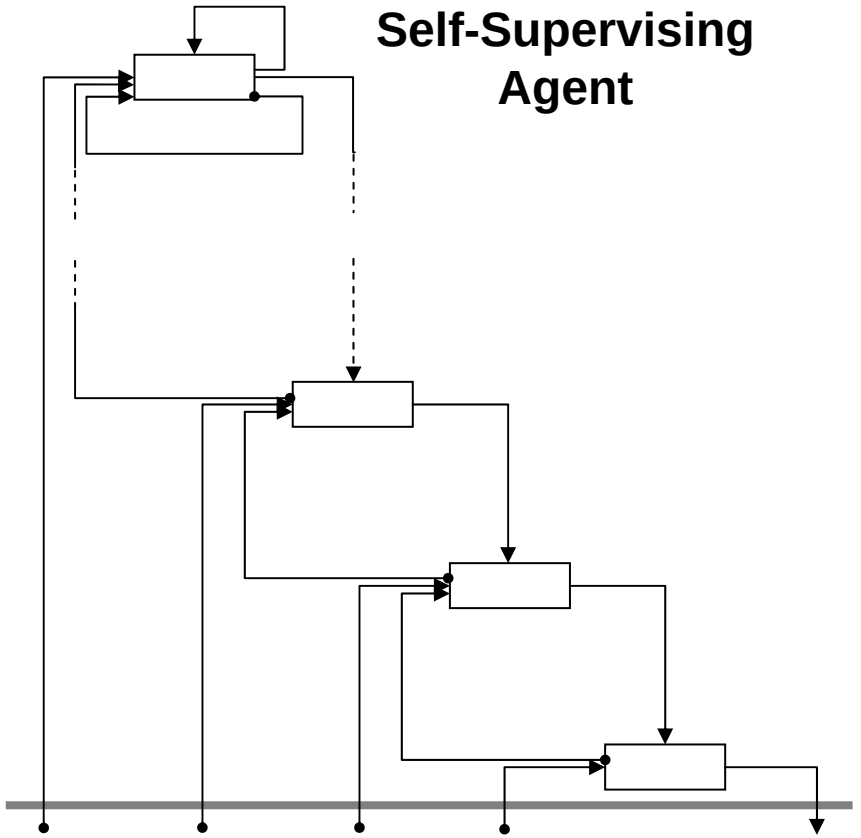
Supervision Chains

- Unique qualities of humans *vs* machines are in the structures for supervision
- Supervisor agents are themselves under supervisory control
  - Supervision Chain



Self-Supervising Agent

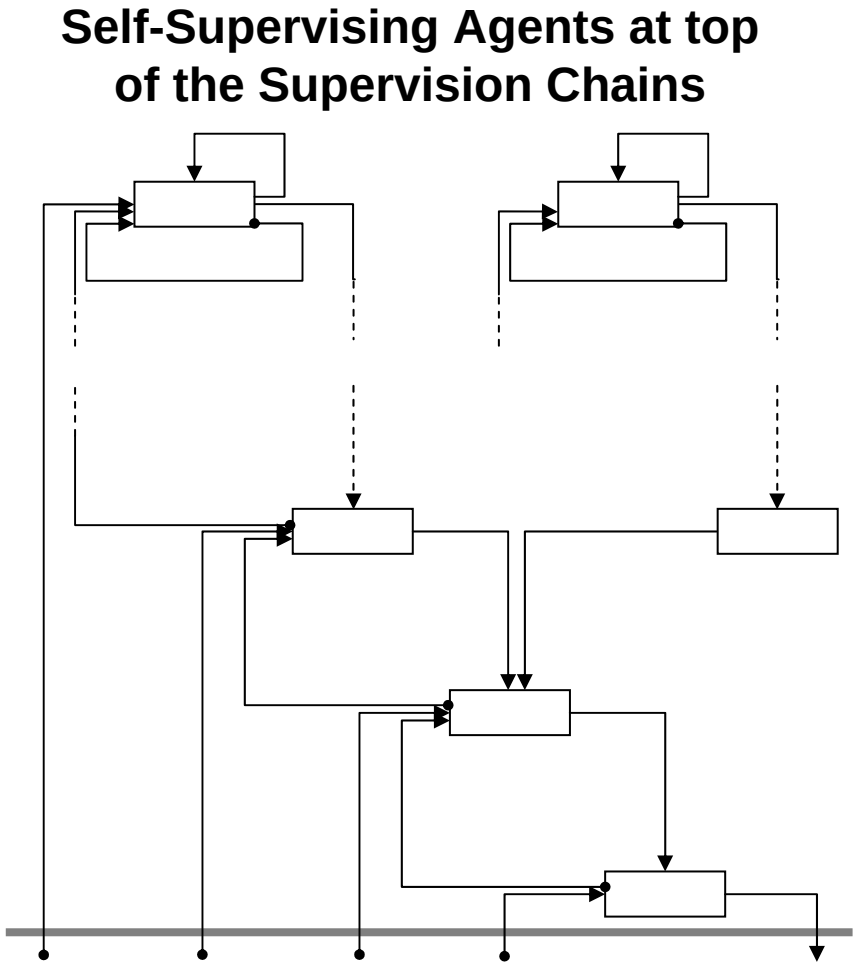
- No unbounded chains
  - The chains must terminate (somehow)
- ***Self-supervising agent*** can perform supervisory control over itself
  - No higher supervisor
- This is ***philosophical autonomy***



Australian Government
Department of Defence
Defence Science and
Technology Organisation

Supervision Chains – General

- Lethal agent has multiple supervisors
  - Supervise at different tempos
- Each supervision chain is capped by a self-supervising agent

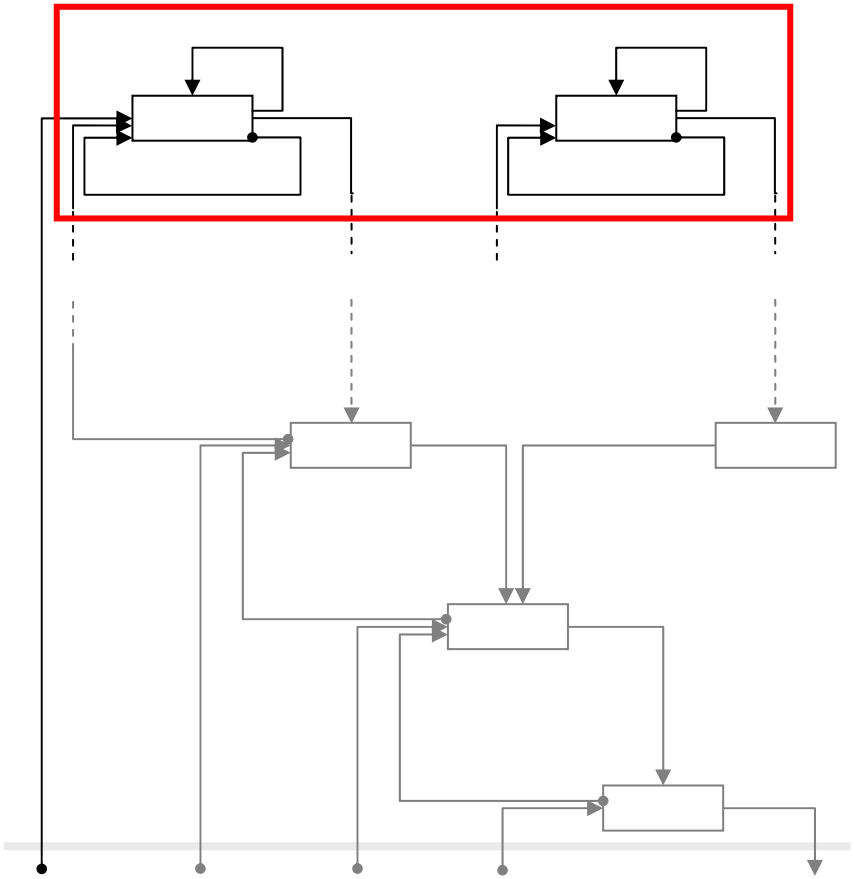


Ethical Agent

- Propose that the ***ethical agent associated with a lethal agent is the self-supervising agent capstoning the supervision chain with the fastest tempo***.

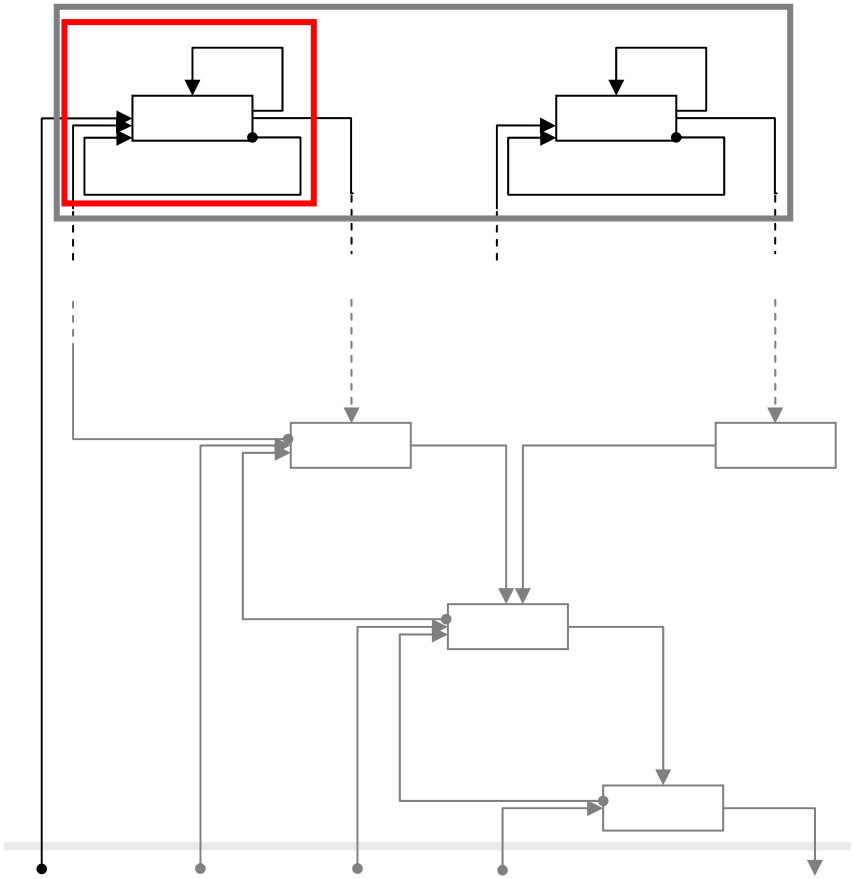
Ethical Agent

- **Capstones a supervision chain**
  - Ethical agent supervises itself, no high supervisor
  - Opposite of “just following orders”
- Corresponds to theory and precedent in war crimes prosecutions



Ethical Agent

- **Fastest tempo**
  - Distinguish between multiple supervision chains
- Builder *vs* User
  - Weapon building is supervisory control at slow tempo
  - Weapon use is supervisory control at fast tempo



Ethical Agent – Application

- Propose that the ***ethical agent associated with a lethal agent is the self-supervising agent capstoning the supervision chain with the fastest tempo***.
- *Application*: For any given wartime casualty, we can “assign responsibility” (identify the ethical agent) by identifying the lethal agent, tracing the supervision chains, and applying the criteria.

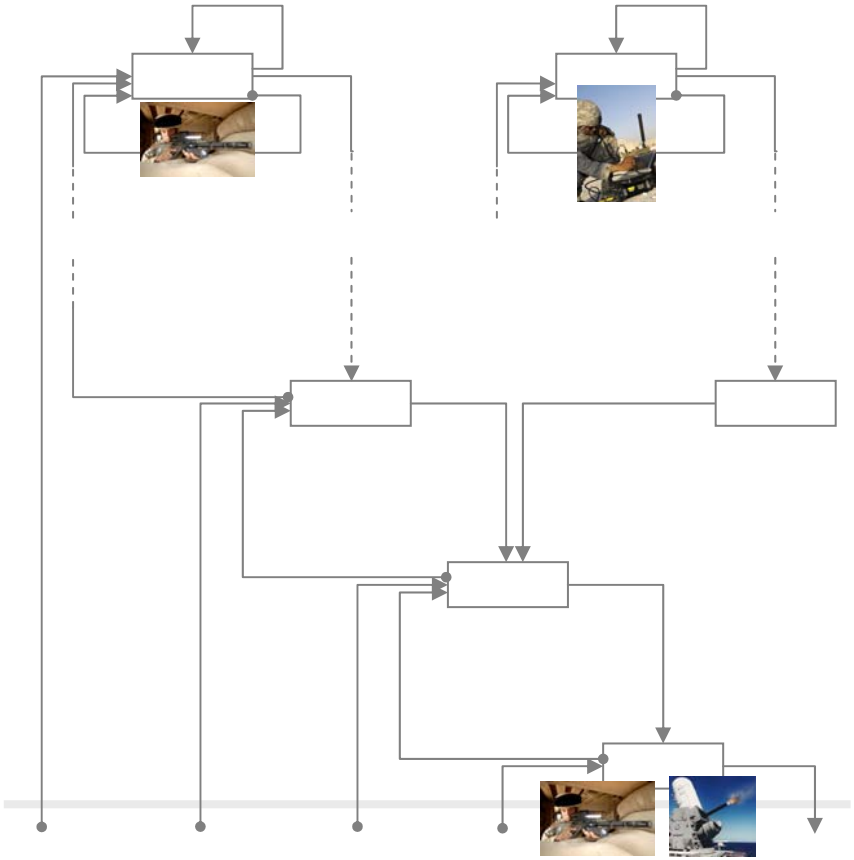
Australian Government
Department of Defence
Defence Science and
Technology Organisation

Feasible Implementations

- Self-supervision is unique to humans
  - … with current tech

therefore

- Ethical agency needs a human being
  - We ***need*** a human "on" the firing loop
  - "in" the loop is ***not sufficient***



Implications for C^2

- **Ethical agency ought to be central in C^2 design** for battle management systems
  - Support humans to be “on” the firing loop (Supervisory control over lethal agent)

  and

  - Support humans to be “on” themselves (Self-supervisory control)
- Autonomy of robotic lethal agents needs to be **matched to ethical agent tempo**
  - Increased autonomy must not compromise capacity for human to be “on” the robot

Conclusions

- Developed a C^2 design for assigning ***ethical agency*** over killing in war
  - Integrated engineering *vs* philosophical notions of “autonomy”
- Established that for foreseeable technology
  - **Necessary** for a human to be “on” the firing loop
  - **Neither necessary nor sufficient** for a human to be “in” the firing loop
- Robotics & automation can be matched into Western military ethics via C^2 design