Ethics and Authority Sharing for Autonomous Armed Robots

RDA2

ECAI 2012 workshop on Right and Duties of Autonomous Agents

Florian Gros, ONERA

Catherine Tessier, ONERA

Thierry Pichevin, Centre de recherche des Ecoles de Saint-Cyr Coëtquidan

ONERA

THE FRENCH AEROSPACE LAB

Preliminary notes

- Robot = 'Autonomous' armed robot

- Difference between :
  - Morality : rules for action, good/evil evaluation
  - Ethics : **reasoning** in case of a conflict or an absence of rules

Introduction

- Increasing use of 'autonomous' robots in numerous domains
- 'Autonomous' robots are supervised by human operators : authority is shared

- Our goal : to consider several ethical issues raised by the deployment of robots in the framework of authority sharing between a robot and a human operator

Authority sharing

- Literature on robot autonomy => omission of the operator or operator considered as a last resort
- Authority => robot and operator equally taken in account as agents [Tessier & Dehais, 2012]
- Agents can have authority over a resource (weapon, etc.)
- Authority conflict : unexpected / misunderstood authority changes [Pizziol, Tessier & Dehais, 2012, this afternoon]

- Authority sharing = **relationship** between agents

Our approach

- Review ethical questions concerning robots
- Consider those questions in the framework of authority sharing
- Study authority conflicts related to ethical issues through :
  - Experimental approach
  - Scenarios

Ethical questions concerning robots - Autonomy

- Kant : Categoric imperative and human autonomy of end
- Rousseau / Rawls : Contract theory

- Operational definition : decisional autonomy of means [Schreckenghost et al., 1998 ; Huang et al., 2005]

- Desirability of fully autonomous robots ?

Ethical questions concerning robots - Responsibility

- Many different approaches
- Causal responsibility *vs.* **Moral responsibility** (Choice)
- Possible leads :
  - Reduced responsibility (negligence, vicarious liability, slave morality) [Lin et al., 2008]
  - Treatment [Lokhorst & Van den Hoven, 2012]
  - Moral status [Abney, 2012 ; Himma, 2007]

Ethical questions concerning robots – Moral status and consciousness

- Moral status : **attributed** to conscious beings
- Two non-discrimination principles [Bostrom & Yudkowsky, 2011] :
  - Principle of Substrate Non-Discrimination
  - Principle of Ontogeny Non-Discrimination

- Triage Turing Test [Sparrow, 2004]

Ethical questions concerning robots – Ethical reasoning

Three different approaches :

- Top-down [Ganascia, 2007; Bringsjord & Taylor, 2012]
- Bottom-up [Lang, 2002; Harms, 2000]
- Hybrid [Arkin, 2007, 2009; Wallach & Allen, 2009; Anderson et al., 2006]

Ethical questions concerning robots – Ethical reasoning

Top-down

- Ethical theory => Set of implementable rules (consequentialism, logic-based)

  + : global, fixed, easily understood rules

  - : frozen, incomplete rules

Bottom-up

- Development of rules and ethical abilites through learning

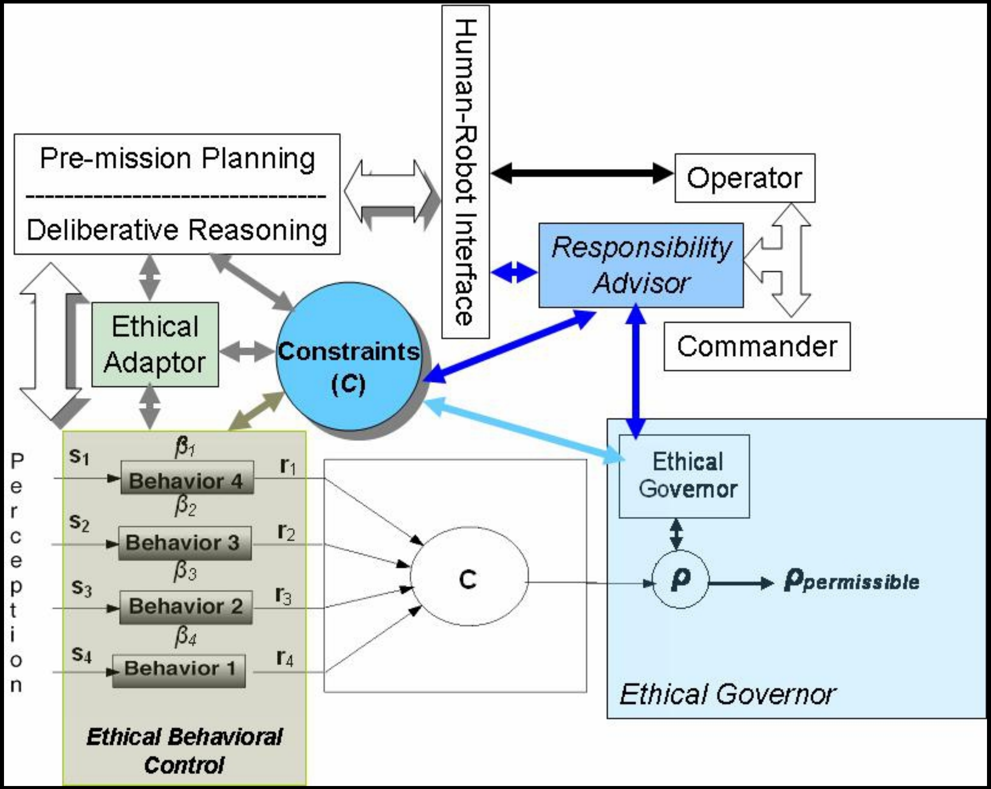
  + : adaptability, optimization

  - : expensive, untraceable, determining a criterion

Ethical questions concerning robots – Ethical reasoning

Hybrid

- Combination of top-down and bottom-up approaches
- Most applicable results
- Three directions :
  - Case-based reasoning [McLaren, 2006 ; Anderson et al., 2006]
  - Virtue ethics [Wallach & Allen, 2009]
  - Arkin's deliberative / reactive architecture [Arkin, 2007]



Ethics and authority sharing

- Reminder : Authority sharing => Relationship between agents

- Autonomy : more decision-making power through authority taking
- Responsibility :
  - Authority to the operator : robot as a tool, responsibility of the operator
  - Authority to the robot : treatment, responsibility of the deployer
- Contract theory => Specific clauses for agents to respect

Ethics and authority sharing

- Moral status and consciousness : better situational assessment on the robot's side through human operator 'state' assessment [Regis et al., 2012 ; Pizziol, Dehais & Tessier, 2011]

- On-going work :
  - Ethical reasoning : assistance by the robot in case of ethical conflict
  - Integration of authority sharing to Arkin's architecture (action evaluation through ethical governor)

Scenarios

- Goal : to test the robot's compliance with a set of rules of engagement during an authority conflict
- Two scenarios designed to simulate a battlefield

- Morally difficult situations (hostile crowd, explosive planting)
- Production of a morally incorrect behaviour => Robot takes authority => Authority conflict => Solving through correct behaviour

Conclusion / Further work

- Assess whether :
  1) Better performance achieved by a human-robot system : better situation assessment, adaptability, compliance with rules through reasoning and authority sharing
  2) Ethical autonomous armed robots : possible with authority sharing ?
- Need for an evolution of the legal and philosophical framework