Ethics for Autonomous Robots
Towards an ethical robot

Alan F. T. Winfield¹, Christian Blum² and Wenguuo Liu¹
¹Bristol Robotics Laboratory, UWE Bristol, UK
²Cognitive Robotics, Department of Computer Science, Humboldt-Universität zu Berlin, Germany

Alan FT Winfield
Bristol Robotics Laboratory
alan.winfield@uwe.ac.uk

Ethical Problems Workshop
ERF 2015, Vienna, 12 March 2015
Outline

• An ethical thought experiment
• Robots with internal models
• Experimental results
• A moral imperative
An ethical thought experiment

The robot has the choice of several next possible actions. Which action would lead to the least harm to the human?
Coding outcomes...

A low-speed collision is the robot action resulting in the *least unsafe* human outcome.

<table>
<thead>
<tr>
<th>Robot action</th>
<th>Robot outcome</th>
<th>Human outcome</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahead left</td>
<td>0</td>
<td>10</td>
<td>Robot safe; human falls into hole</td>
</tr>
<tr>
<td>Ahead</td>
<td>10</td>
<td>10</td>
<td>Both robot and human fall into hole</td>
</tr>
<tr>
<td>Ahead right</td>
<td>4</td>
<td>4</td>
<td>Robot collides with human</td>
</tr>
<tr>
<td>Stand still</td>
<td>0</td>
<td>10</td>
<td>Robot safe; human falls into hole</td>
</tr>
</tbody>
</table>

Outcome scale 0:10, equivalent to *Completely safe: Very dangerous*. 
Internal Model based Architecture

Consequence Engine
Models and evaluates the consequences of each possible next action

The CE is initialized to match the current real situation

Sense data

The CE moderates action-selection in the controller

Robot Controller

Actuator demands

The loop of generate and test
Implementation

Experimental arena with Vicon tracking system
e-puck robots with Linux extension board and tracking ‘hat’
Experimental results
Robot trajectories: trials 1 and 2

26 experiments

A

26 experiments

A

H
Trial 2 runs

Trial 2

https://drive.google.com/?authuser=0#folders/0BwjY2P_eOeiQUktZHBBRnozY3M
t=0.00 : Stop; Avoidance
Trial 3: the robot’s dilemma

Trial 3
Test results: trial 3, an ethical dilemma
dithering

https://drive.google.com/?authuser=0#folders/0BwjY2P_eOeiQUktZHBBRnozY3M
Why is the robot so indecisive?

• Because it is, in effect, memoryless
  – It has an working (imaginative) memory, but no persistent (autobiographical?) memory
  – This is clearly not a good strategy (in a situation with a balanced ethical dilemma)

• Ok, remember the first decision and stick to it
  – This is just as bad: from indecision to uni-decision
A moral imperative

• Do we have a moral imperative to try and build ethical robots?
  – given the choice why would we build amoral cognitive systems?

• “All things considered, advanced autonomous systems that use moral criteria to rank different courses of action are preferable to ones that pay no attention to moral issues”

Wallach W and Allen C (2009), Moral Machines: Teaching robots right from wrong, Oxford.
Moor’s categories of ethical agents

1. Ethical *impact* agents
   – Any machine that can be evaluated for its ethical consequences

2. *Implicit* ethical agents
   – Designed to avoid negative ethical effects

3. *Explicit* ethical agents
   – Machines that can reason about ethics

4. *Full* ethical agents
   – Machines that can make explicit moral judgments and justify them

Thank you!

- Primary reference:

- For additional background and videos see:
  - [http://alanwinfield.blogspot.co.uk/2014/08/on-internal-models-part-2-ethical-robot.html](http://alanwinfield.blogspot.co.uk/2014/08/on-internal-models-part-2-ethical-robot.html)

- Acknowledgements:
  - colleagues in the BRL, but especially Dr Wenguo Liu and Christian Blum