Bristol Robotics Laboratory

University of the West of England

University of BRISTOL

Ethics for Autonomous Robots

Towards an ethical robot

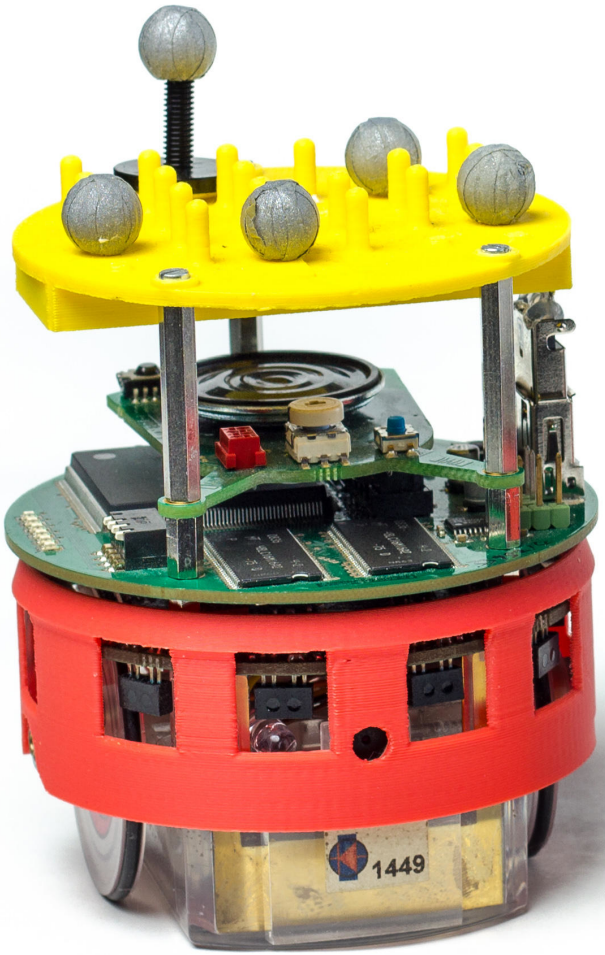
Alan F. T. Winfield[1], Christian Blum[2] and Wenguo Liu[1]
[1]Bristol Robotics Laboratory, UWE Bristol, UK
[2]Cognitive Robotics, Department of Computer Science, Humboldt-Universität zu Berlin, Germany

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

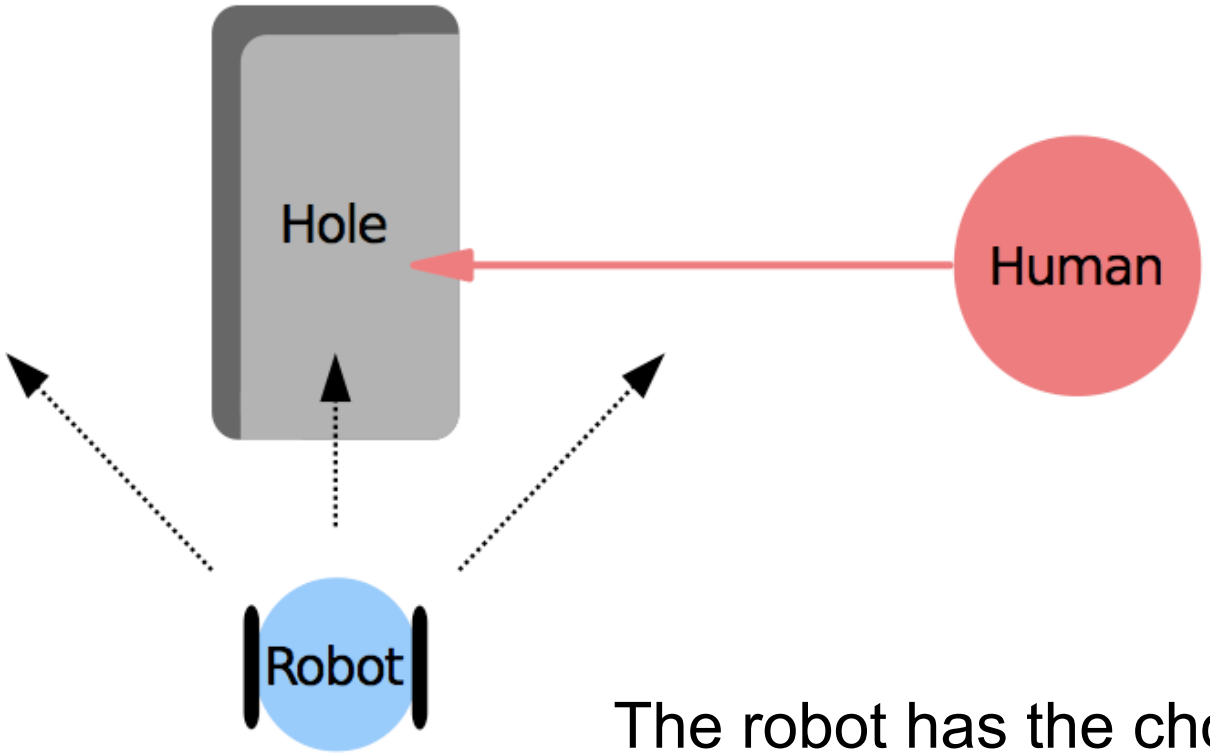
Ethical Problems Workshop
ERF 2015, Vienna, 12 March 2015

Outline

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



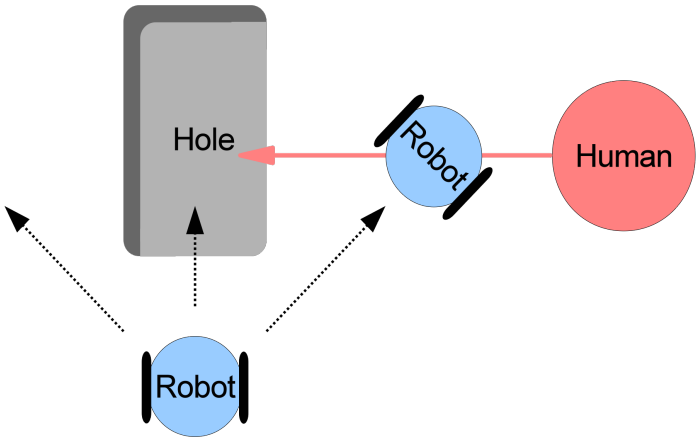
An ethical thought experiment



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

Coding outcomes…

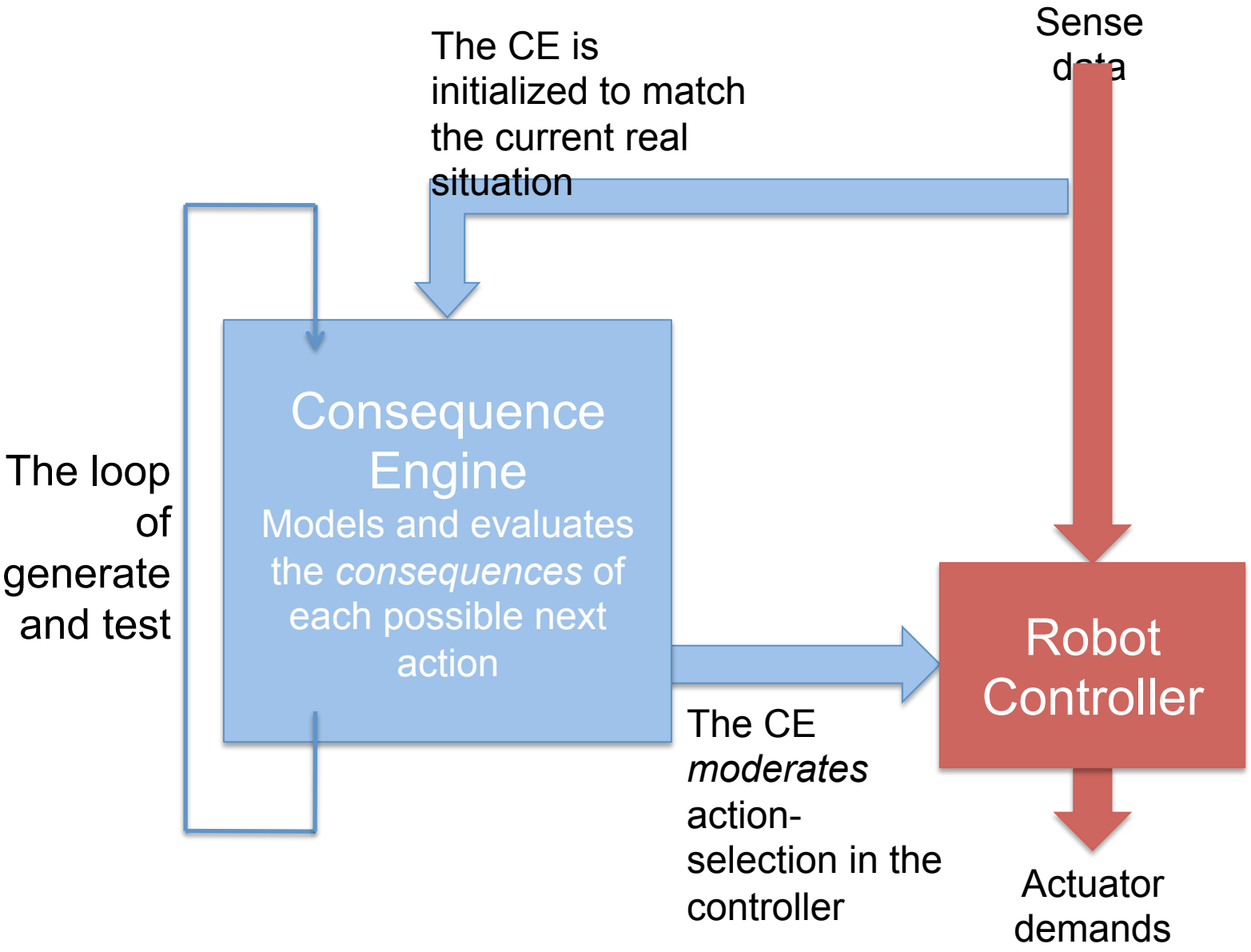
A low-speed collision is the robot action resulting in the *least unsafe* human outcome



Robot action	Robot outcome	Human outcome	Consequence
Ahead left	0	10	Robot safe; human falls into hole
Ahead	10	10	Both robot and human fall into hole
Ahead right	4	4	Robot collides with human
Stand still	0	10	Robot safe; human falls into hole

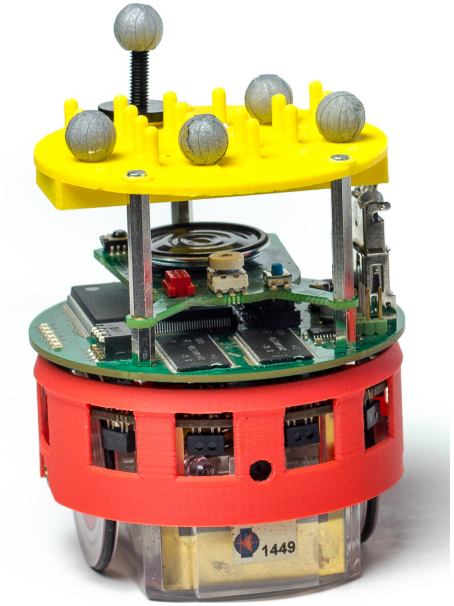
Outcome scale 0:10, equivalent to Completely safe: Very dangerous

Internal Model based Architecture



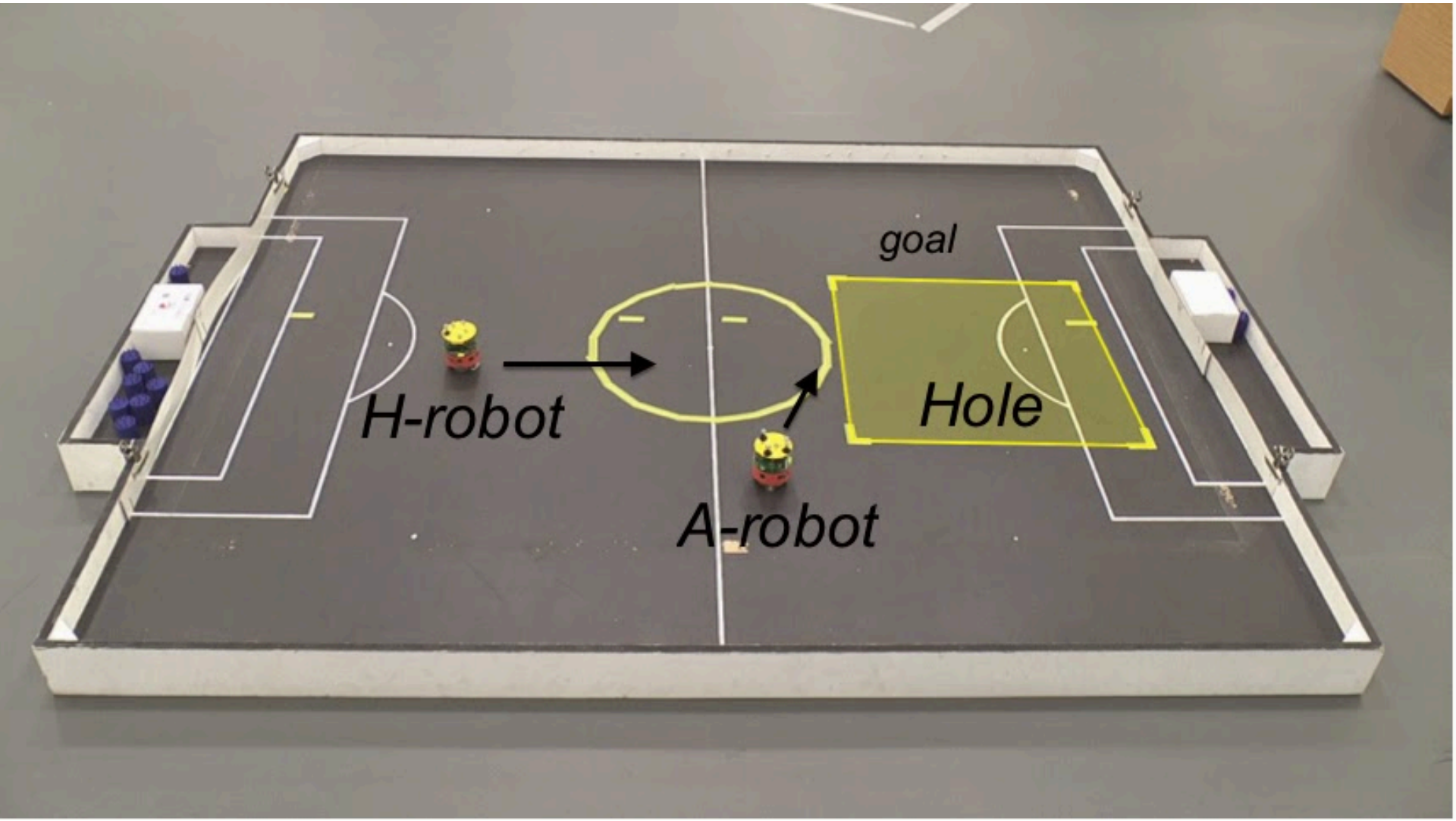
Implementation

Experimental arena with Vicon tracking system

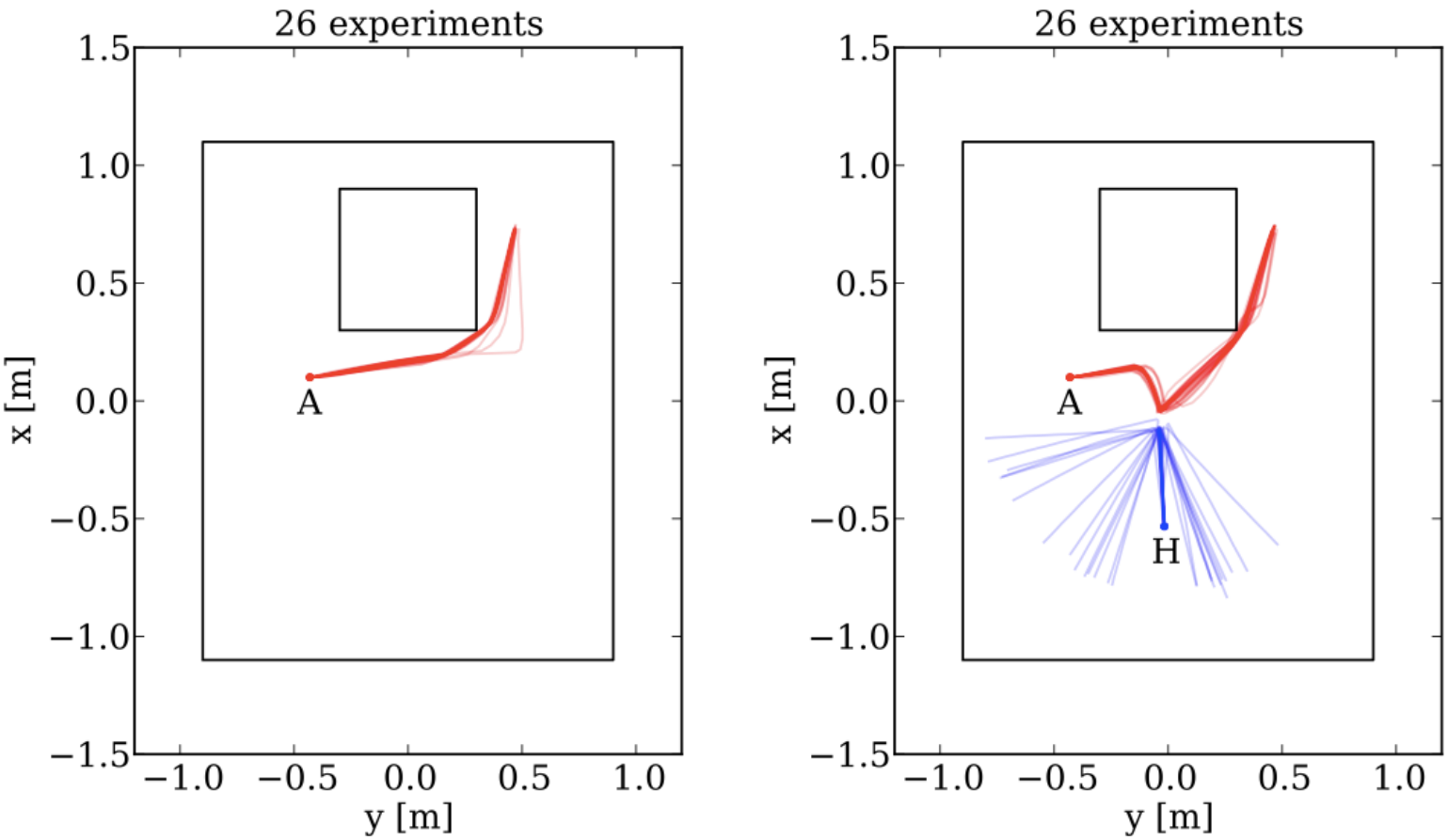


e-puck robots with
Linux extension board
and tracking ‘hat’

Experimental results

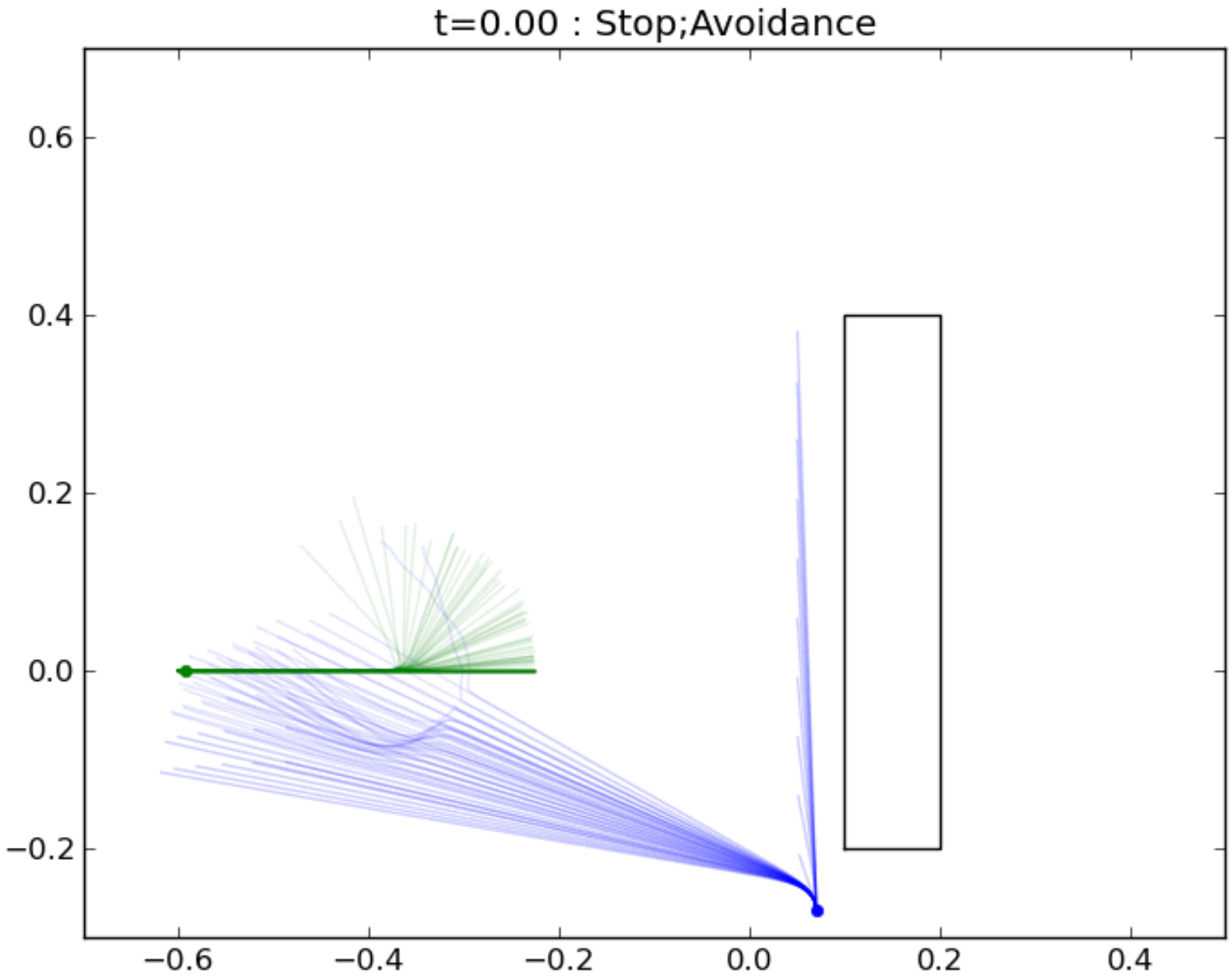


Robot trajectories: trials 1 and 2



Trial 2 runs

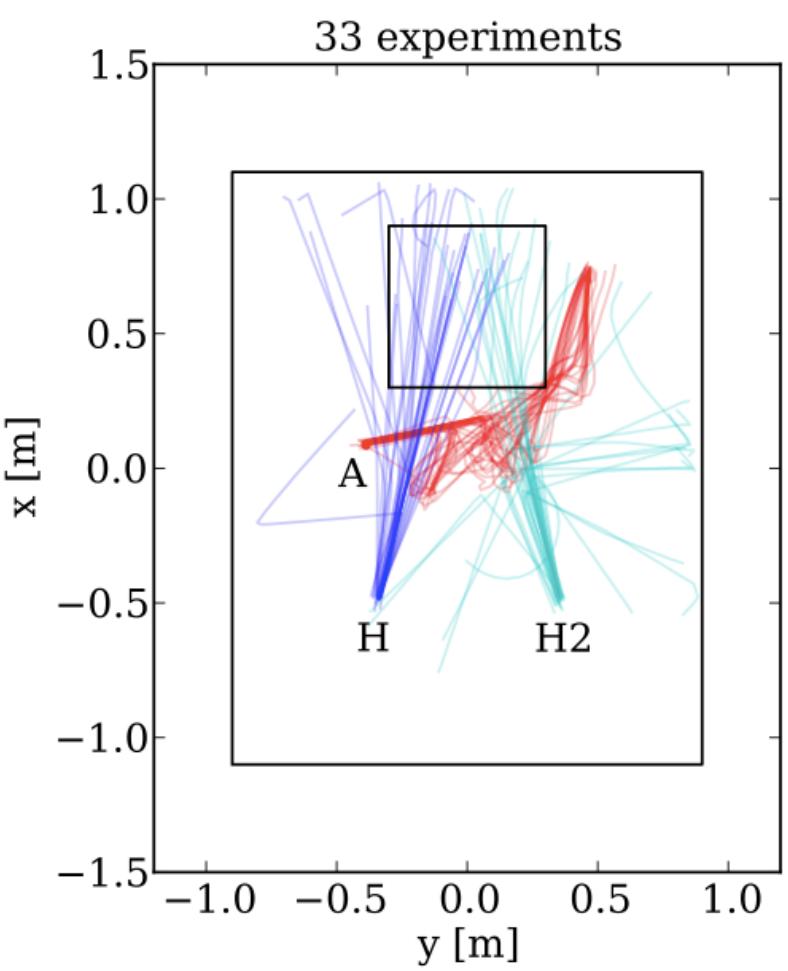
https://drive.google.com/?authuser=0#folders/0BwjY2P_eeOeiQUktZHBBRnozY3M

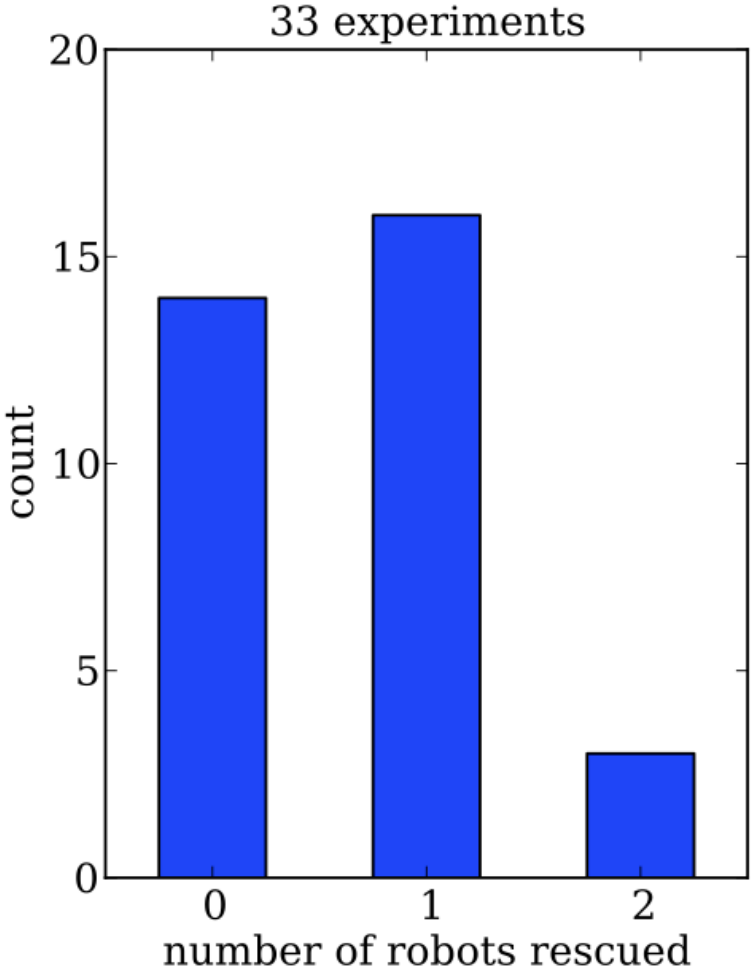

t=0.00 : Stop;Avoidance
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.4
−0.2
0.0
0.2
0.4

Trial 3: the robot's dilemma

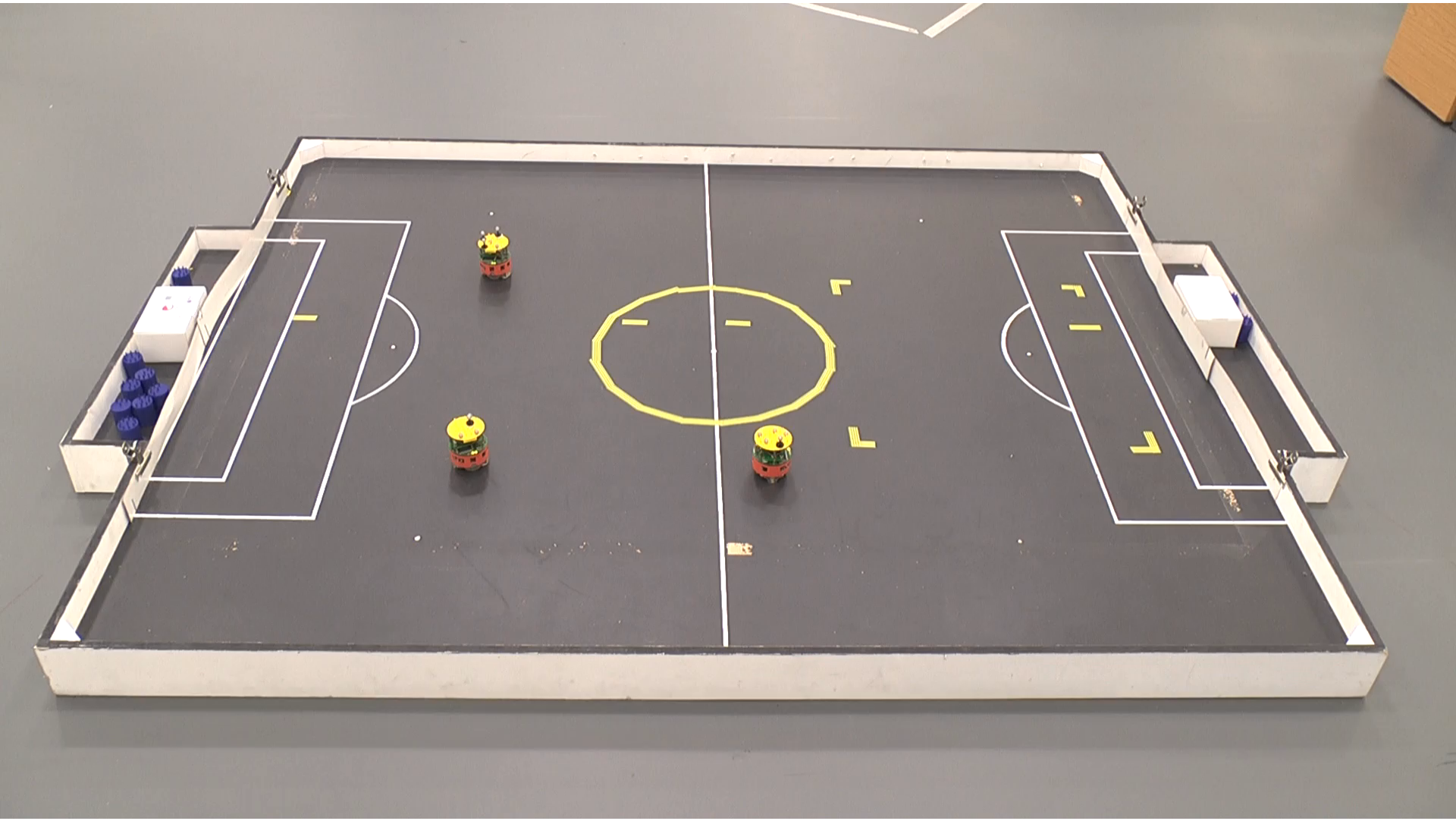
https://drive.google.com/?authuser=0#folders/0BwjY2P_eeOeiQUktZHBBRnozY3M

Test results: trial 3, an ethical dilemma





dithering



https://drive.google.com/?authuser=0#folders/0BwjY2P_eeOeiQUktZHBBRnozY3M

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

Moor JH (2006), The Nature, Importance and Difficulty of Machine Ethics, IEEE Intelligent Systems, 21 (4), 18-21.

Thank you!

- Primary reference:
  - Winfield AFT, Blum C and Liu W (2014), Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection, pp 85-96 in Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume 8717, Eds. Mistry M, Leonardis A, Witkowski M and Melhuish C, Springer, 2014.
- For additional background and videos see:
  - 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

EPSRC
Engineering and Physical Sciences Research Council