Bristol Robotics Laboratory

University of the West of England, BRISTOL

University of BRISTOL

From robot swarms to ethical robots: the challenges of verification and validation part 1

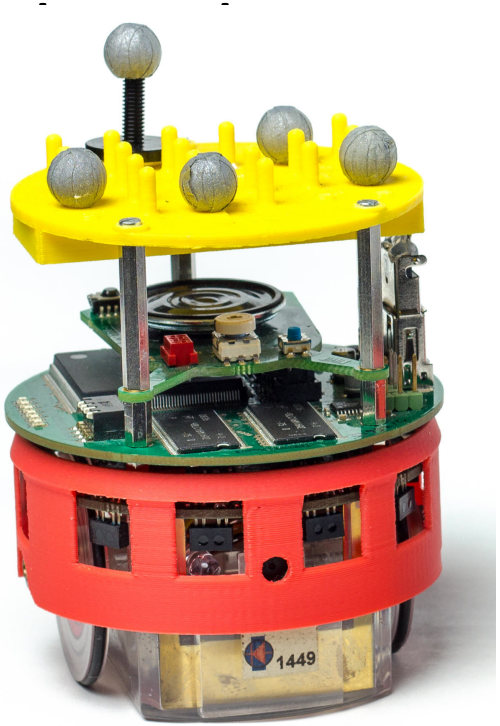
Robots with Internal Models

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

RoboCheck Winter School,
University of York
1 Dec 2015

Outline

- Internal Models
- A Generic Architecture for situational imagination
- Implementation and experiments w robots:
  - Towards an Ethical Robot
  - The Corridor Experiment



brl
Bristol Robotics Laboratory

Consider the internal model

- It is an internal mechanism for representing both the system itself *and* its environment
  - example: a robot with a *simulation* of itself *and* its currently perceived environment, *inside itself*
- The mechanism might be centralized, distributed, or emergent

"..an internal model allows a system to look ahead to the future consequences of current actions, without actually committing itself to those actions"

John Holland (1992), Complex Adaptive Systems, Daedalus.

Using internal models

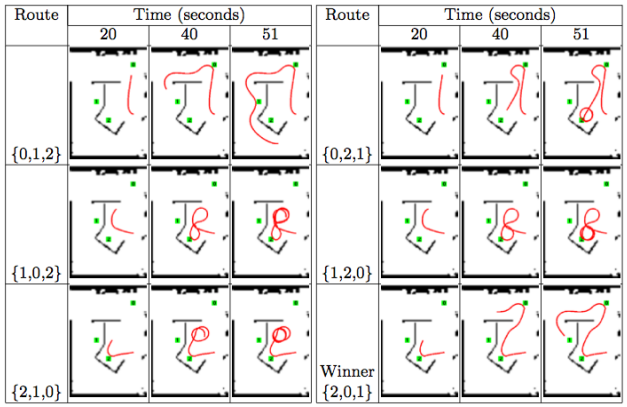
- *Internal models* can provide a minimal level of *functional self-awareness*
  - sufficient to allow complex systems to ask *what-if* questions about the consequences of their next possible actions, for safety
- Following Dennett* an internal model can generate and test *what-if* hypotheses:
  - *what if* I carry out action *x*..?
  - of several possible next actions x_i, *which* should I choose?

brl Bristol Robotics Laboratory

*Dennett, D. (1995). Darwin's Dangerous Idea, London, Penguin.

Examples 1

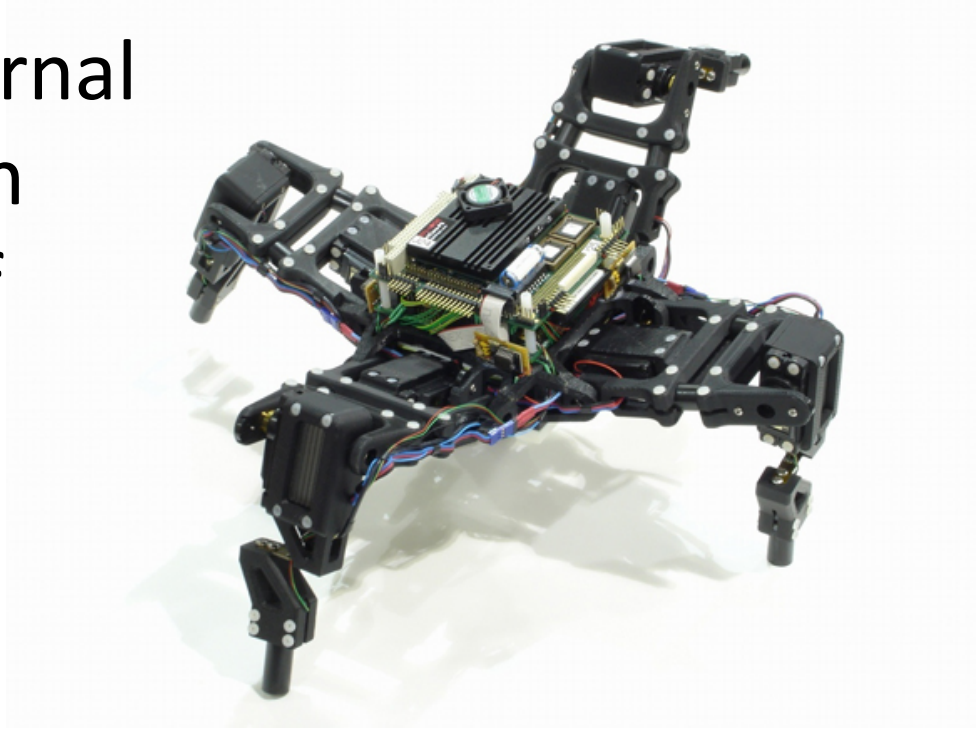
- A robot using self-simulation to plan a safe route with incomplete knowledge



Vaughan, R. T. and Zuluaga, M. (2006). Use your illusion: Sensorimotor self- simulation allows complex agents to plan with incomplete self-knowledge, in Proceedings of the International Conference on Simulation of Adaptive Behaviour (SAB), pp. 298–309.

Examples 2

- A robot with an internal model that can learn how to control itself

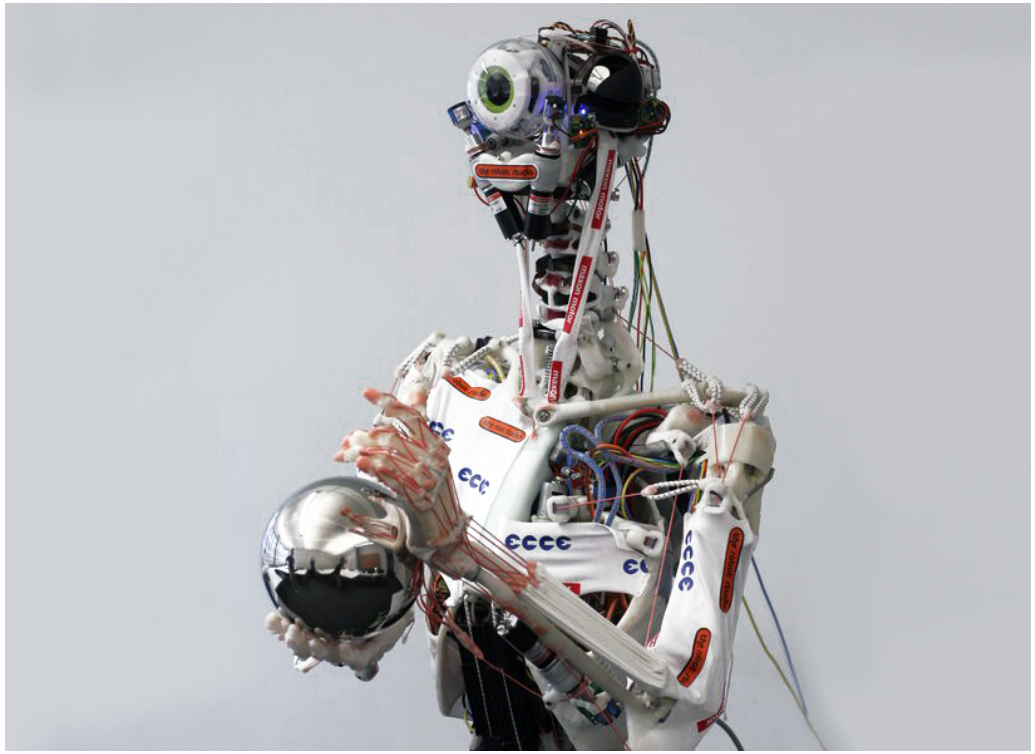


Bongard, J., Zykov, V., Lipson, H. (2006) Resilient machines through continuous self-modeling. Science, 314: 1118-1121.

brl
Bristol Robotics Laboratory

Examples 3

- ECCE-Robot
  - A robot with a complex body uses an internal model as a ‘functional imagination’

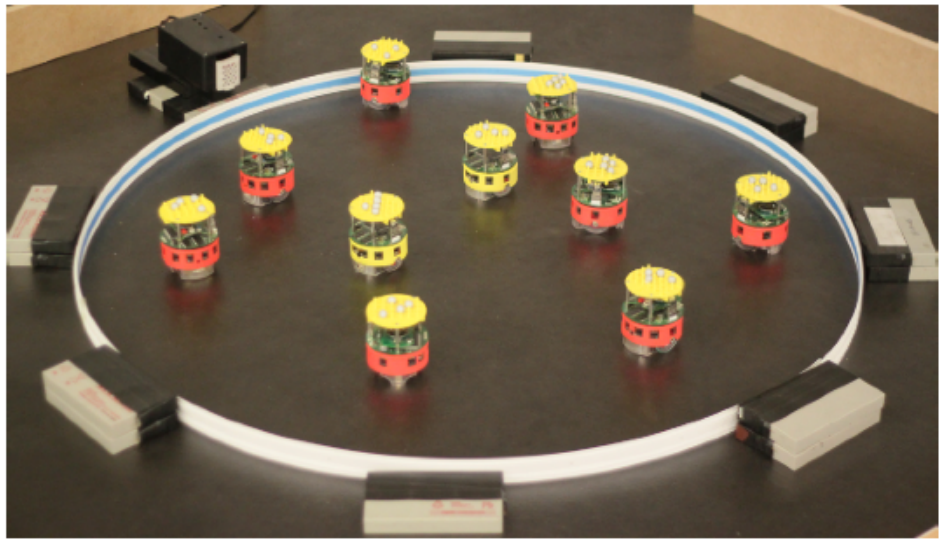


Marques, H. and Holland, O. (2009). Architectures for functional imagination, Neurocomputing 72, 4-6, pp. 743–759.

Diamond, A., Knight, R., Devereux, D. and Holland, O. (2012). Anthropomimetic robots: Concept, construction and modelling, International Journal of Advanced Robotic Systems 9, pp. 1–14.

Examples 4

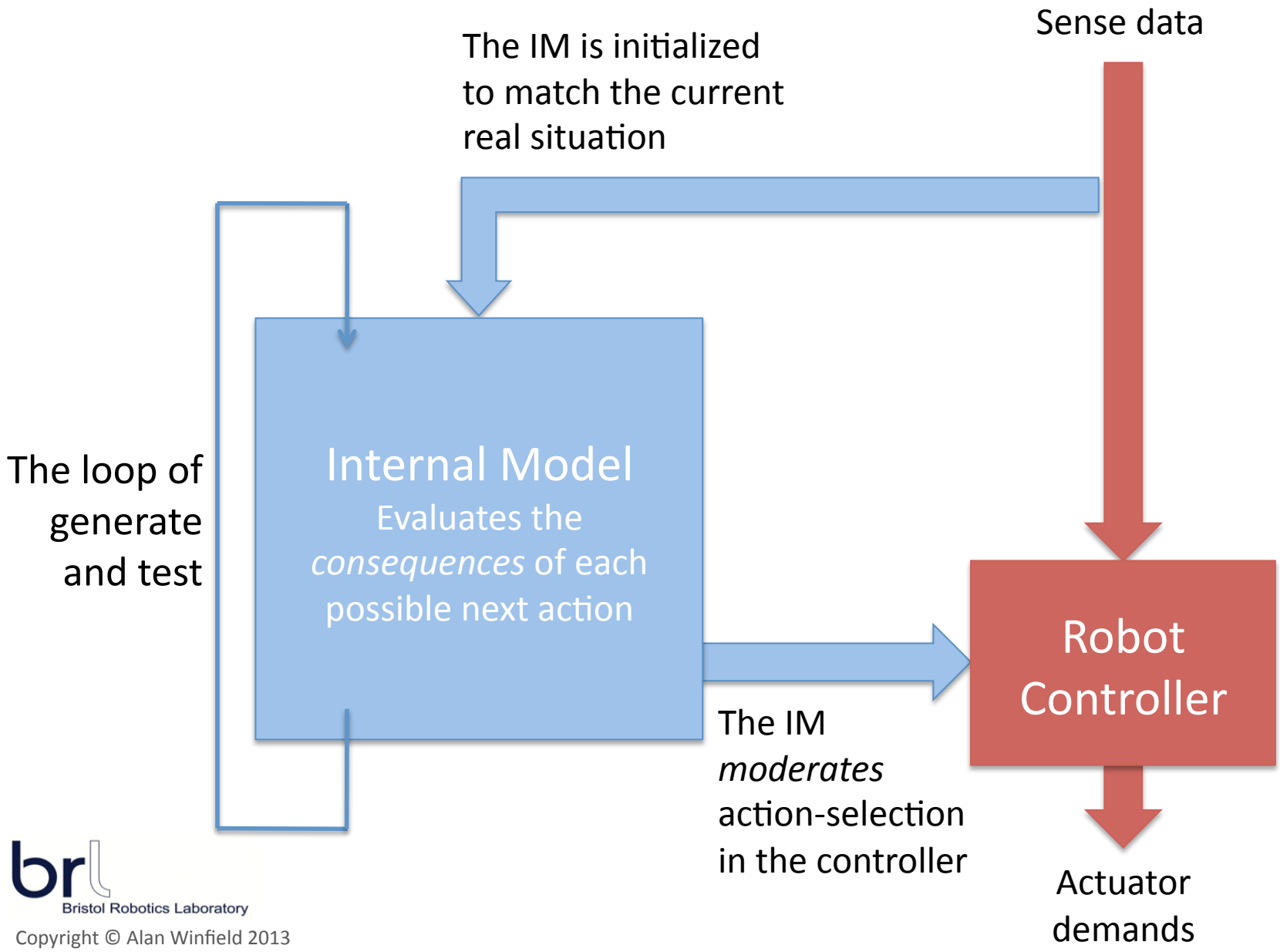
- A distributed system in which each robot has an internal model of itself and the whole system
  - Robot controllers and the internal simulator are *co-evolved*



O'Dowd P, Studley M and Winfield AFT (2014) The distributed co-evolution of an on-board simulator and controller for swarm robot behaviours. Evolutionary Intelligence, 7 (2).

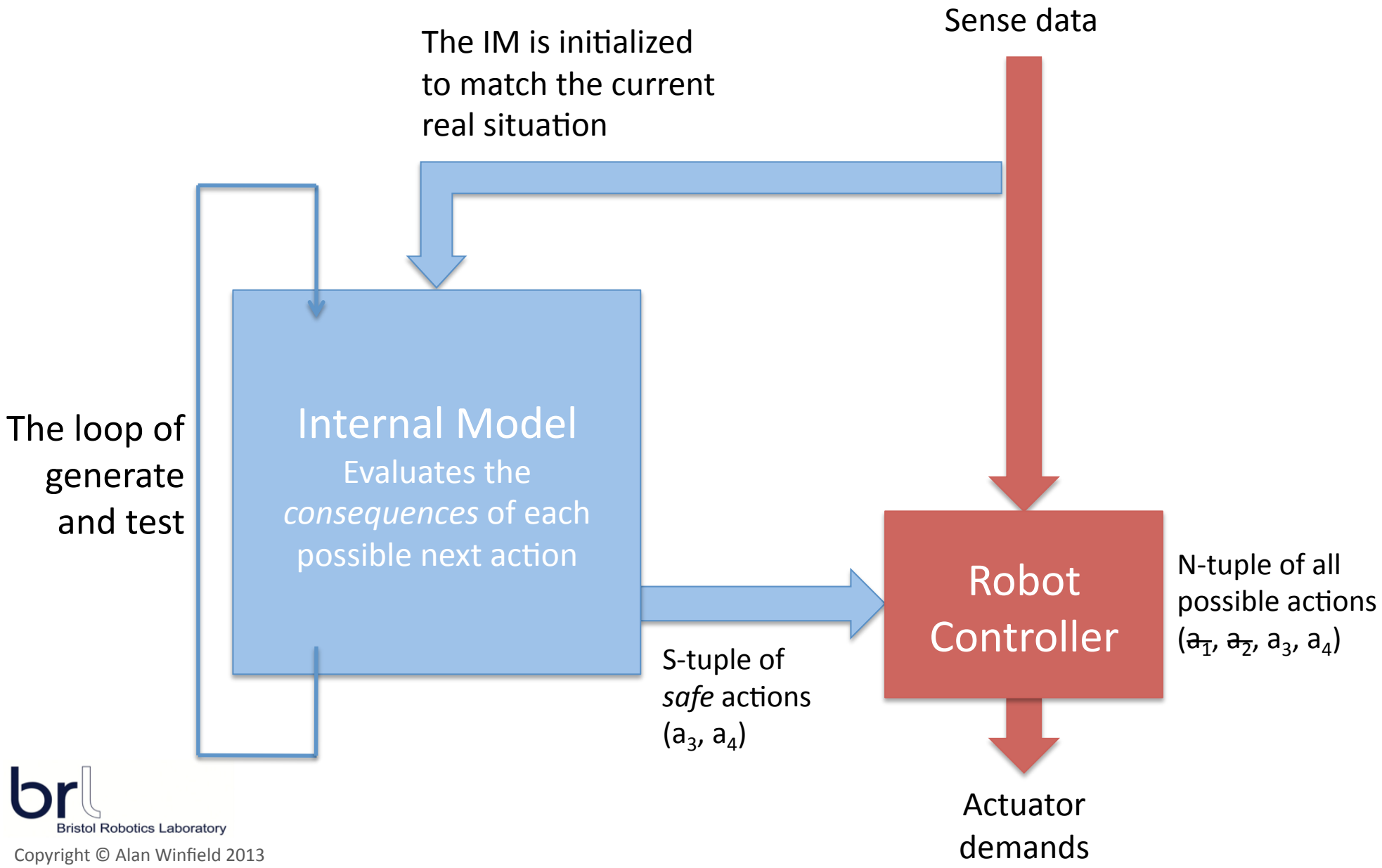
brl
Bristol Robotics Laboratory

A Generic IM Architecture for Safety

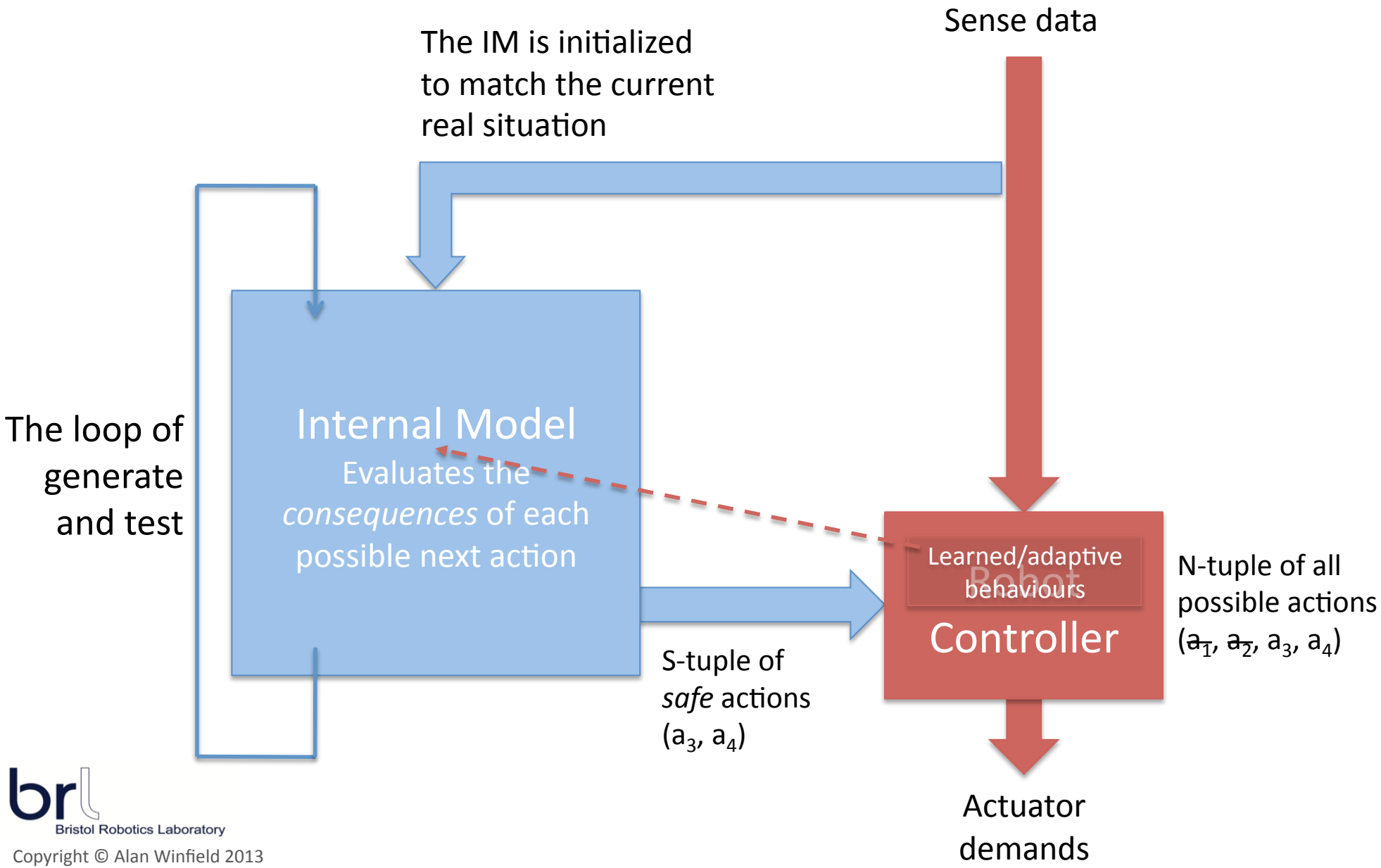


brl
Bristol Robotics Laboratory

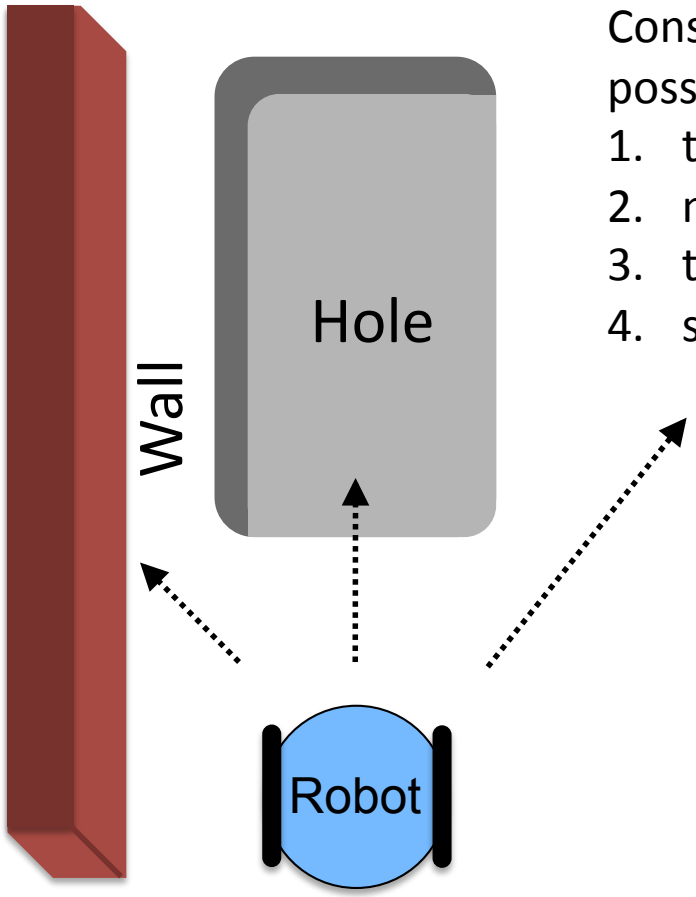
A Generic IM Architecture for Safety



Extending into Adaptivity



A scenario with safety hazards

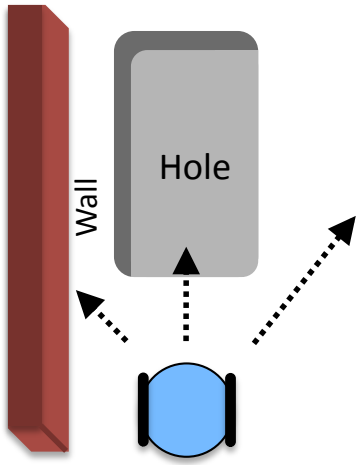


Consider a robot that has four possible next actions:

1. turn left
2. move ahead
3. turn right
4. stand still

Bristol Robotics Laboratory

A scenario with safety hazards



Consider a robot that has four possible next actions:

1. turn left
2. move ahead
3. turn right
4. stand still

Robot action	**Position change**	**Robot outcome**	**Consequence**
Ahead left	5 cm	Collision	Robot collides with wall
Ahead	10 cm	Collision	Robot falls into hole
Ahead right	20 cm	No-collision	Robot safe
Stand still	0 cm	No-collision	Robot safe

brl Bristol Robotics Laboratory

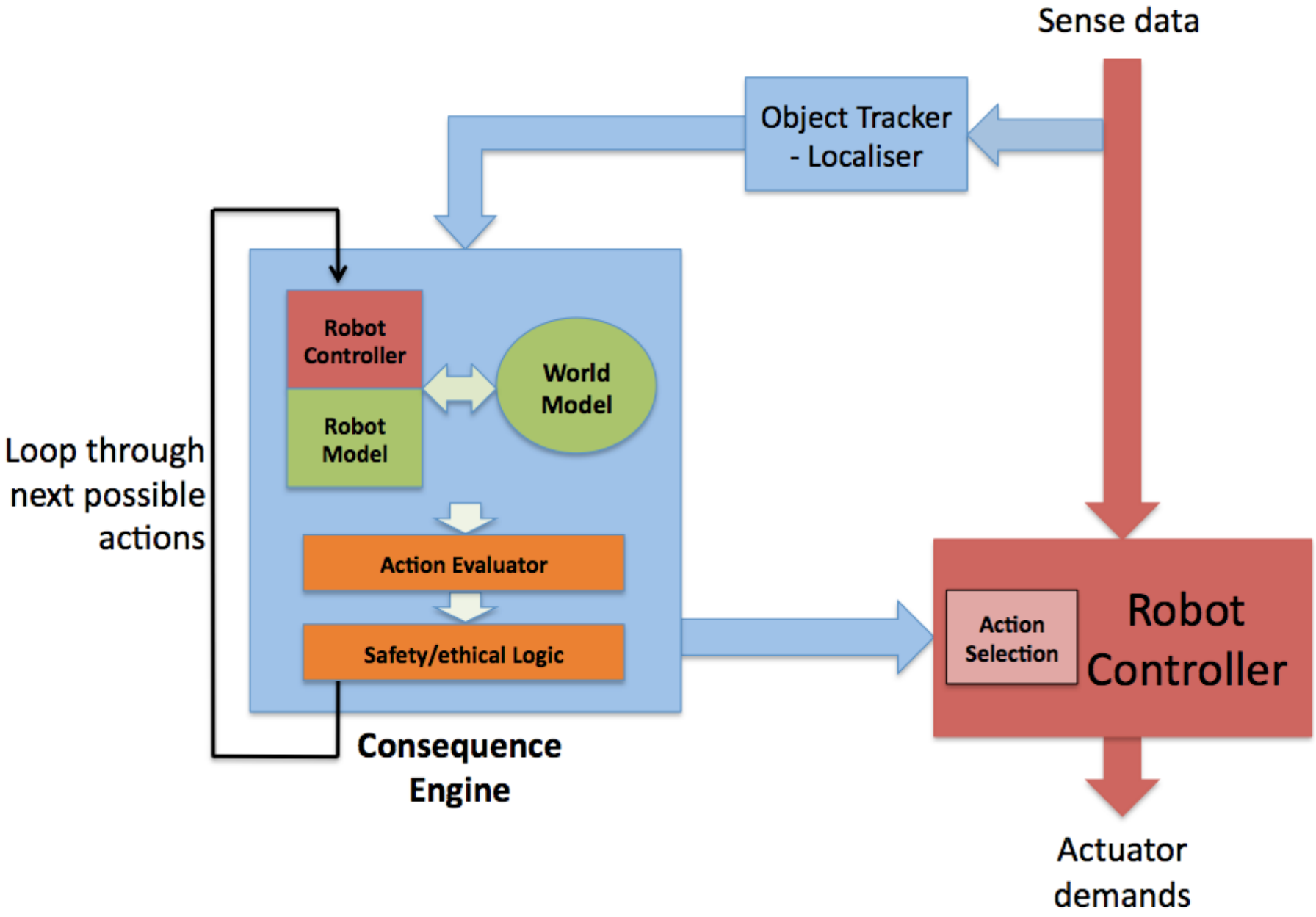
Implementation

Experimental arena with Vicon tracking system

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

brl
Bristol Robotics Laboratory

Real time self- and other-simulation

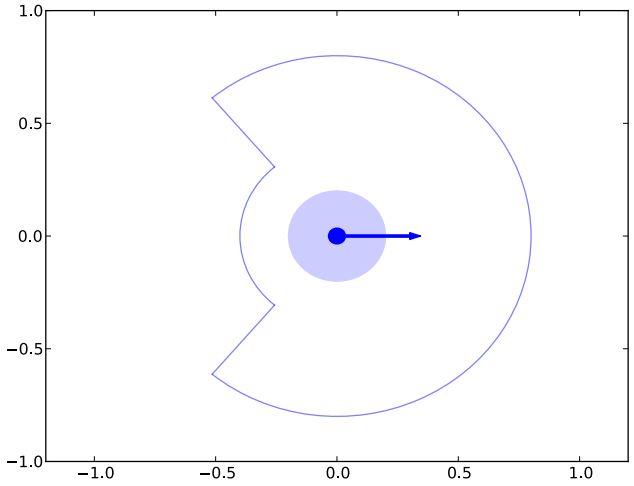


Simulation budget

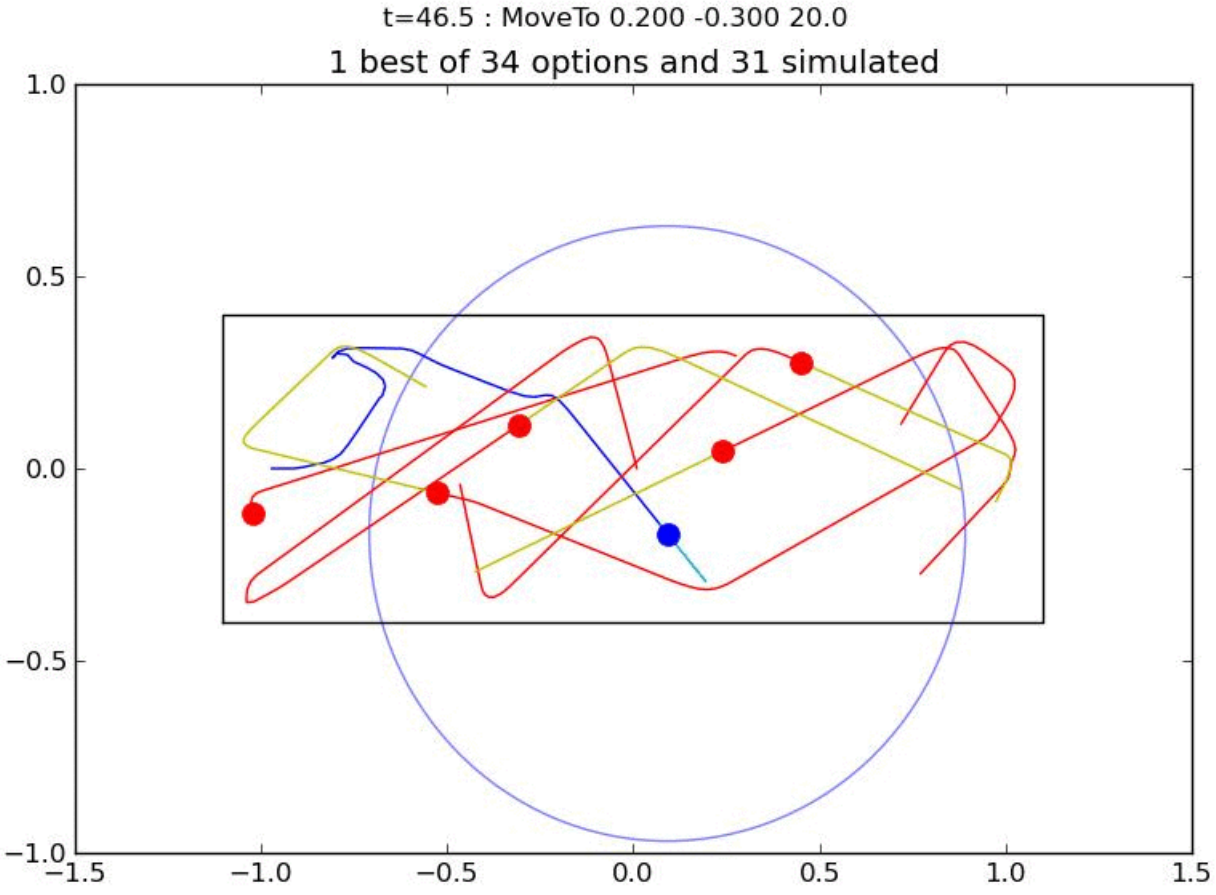
- Internal model uses open source simulator Stage
- Stage runs at about 600 times real time
- Consequence Engine cycles at 2Hz
- 10s, i.e. 0.7m, simulation horizon
- 30 next possible actions

The corridor experiment

- One robot (blue) with self- and other-simulation must negotiate a corridor with five other robots (red)



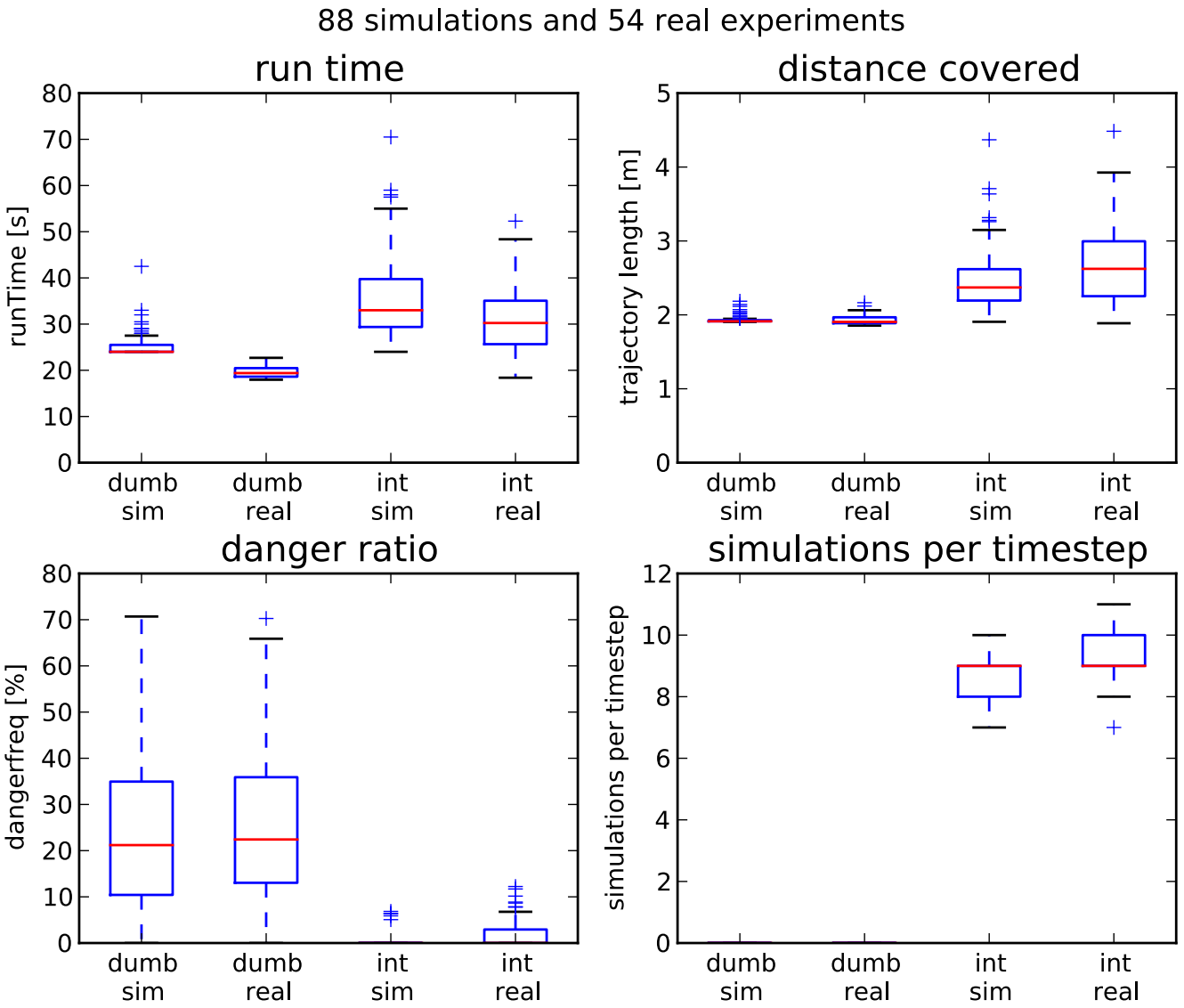
The radius of attention



brl
Bristol Robotics Laboratory

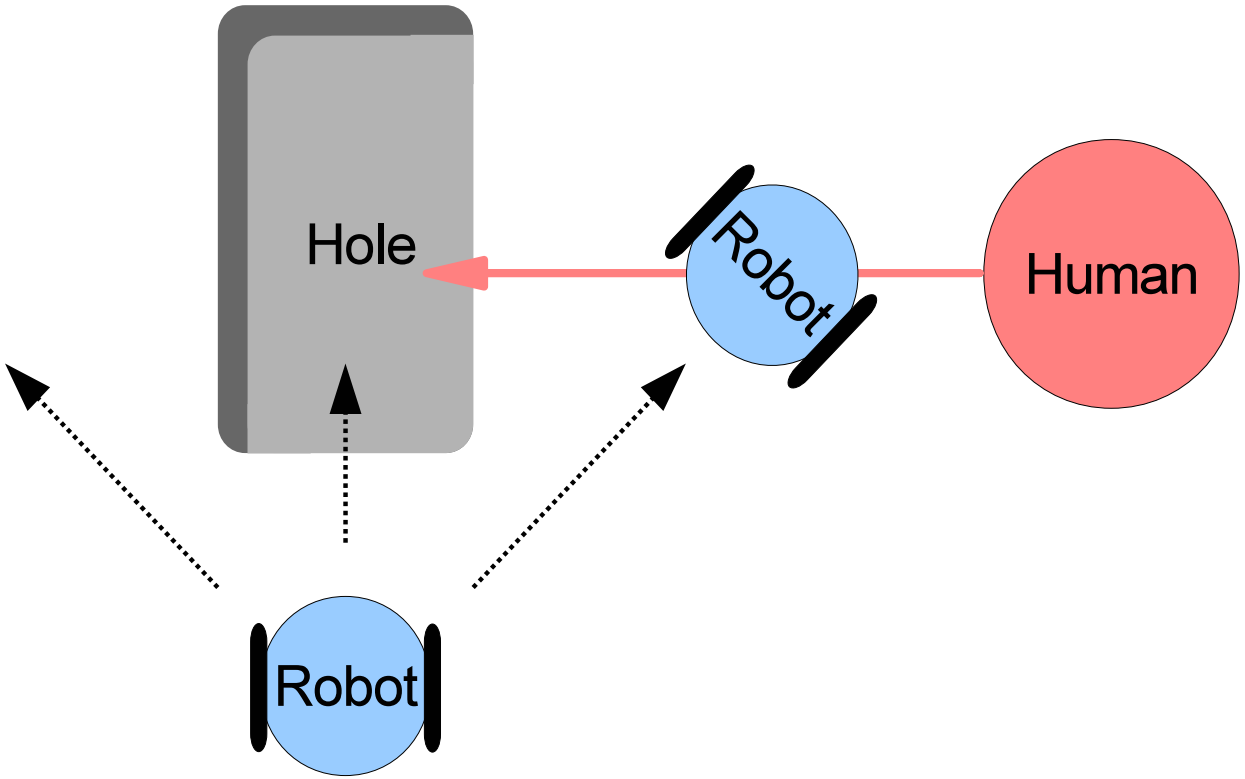
Results – simulated and real robots

comparing simple obstacle avoidance with internal modelling



Bristol Robotics Labo

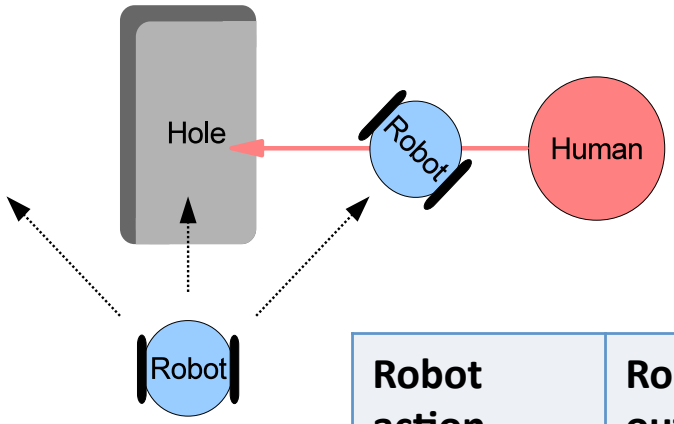
Towards an ethical robot



Which robot action would lead to the least harm to the human?

Bristol Robotics Laboratory

Towards an ethical robot



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

Which robot action would lead to the least harm to the human?

Bristol Robotics Laboratory

Combining safety and ethical rules

```
IF for all robot actions, the human is equally safe
THEN (* default safe actions *)
   output s-tuple of safe actions
ELSE (* ethical actions *)
   output s-tuple of actions for least unsafe human
   outcomes
```

Consider Asimov's 1st and 3rd laws of robotics:

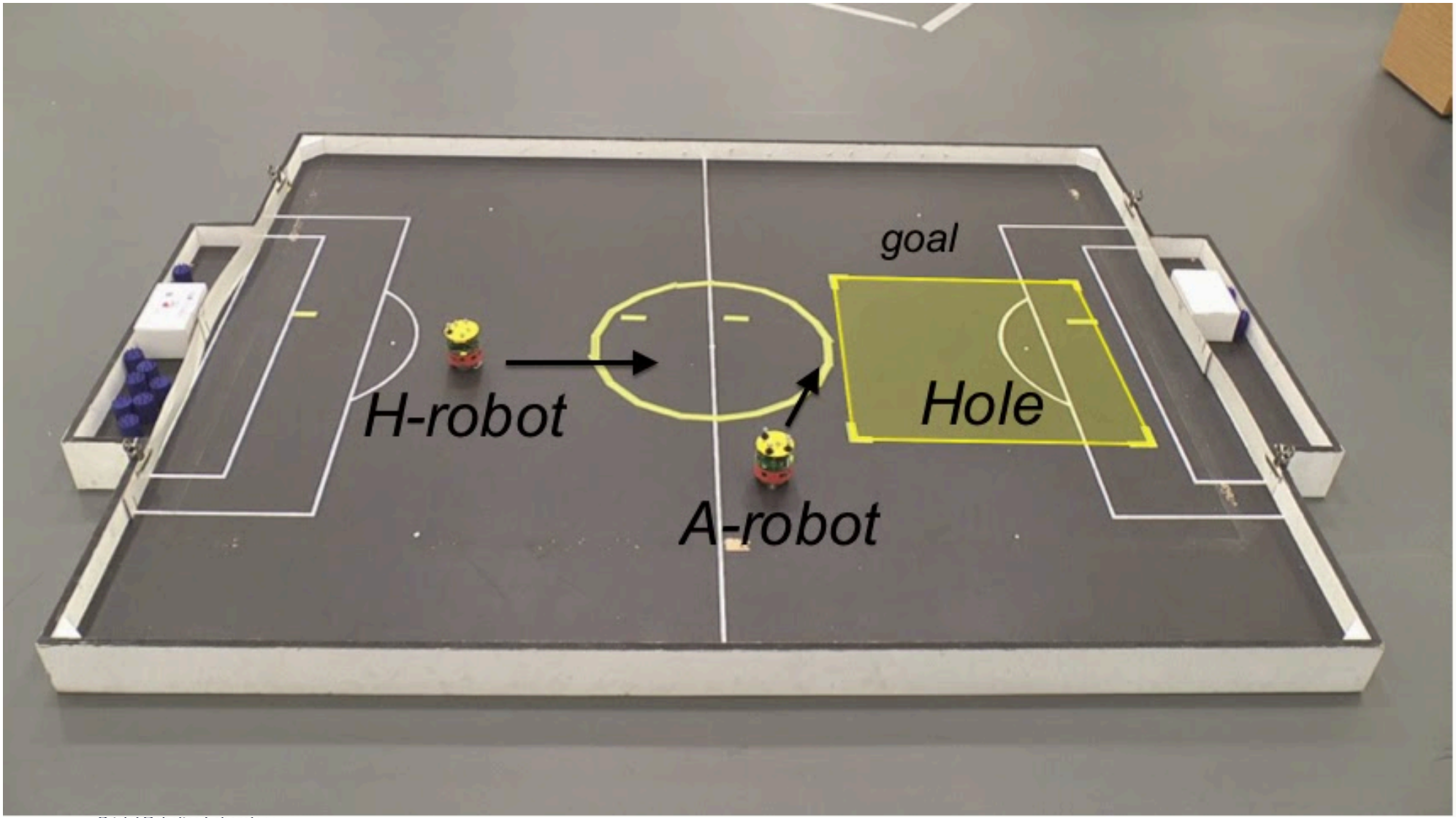
(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm,

(3) A robot must protect its own existence as long as such protection does not conflict with the First (or Second) Laws

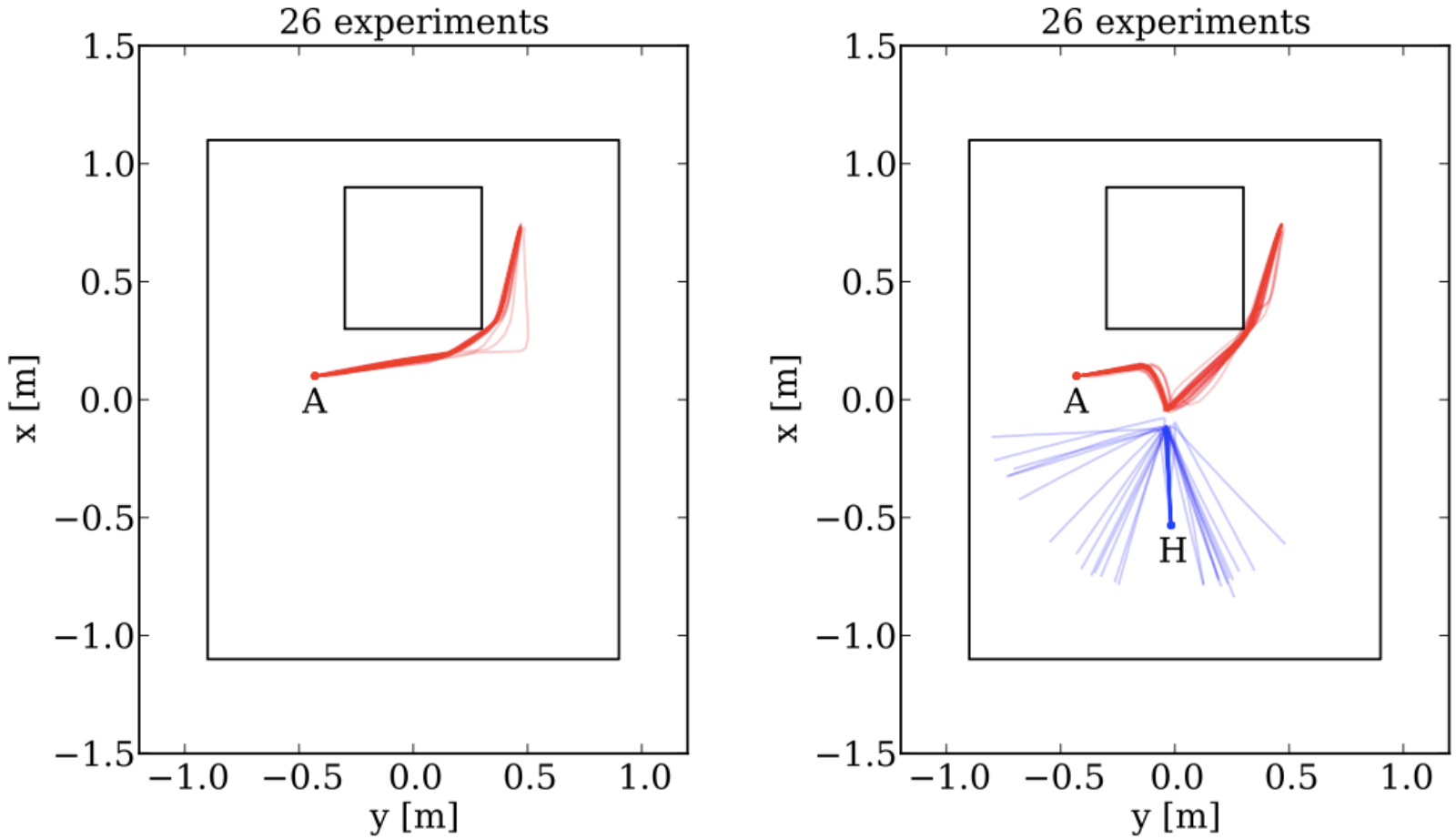
Isaac Asimov, *I, ROBOT*, 1950

brl
Bristol Robotics Laboratory

Experimental results



Robot trajectories: trials 1 and 2

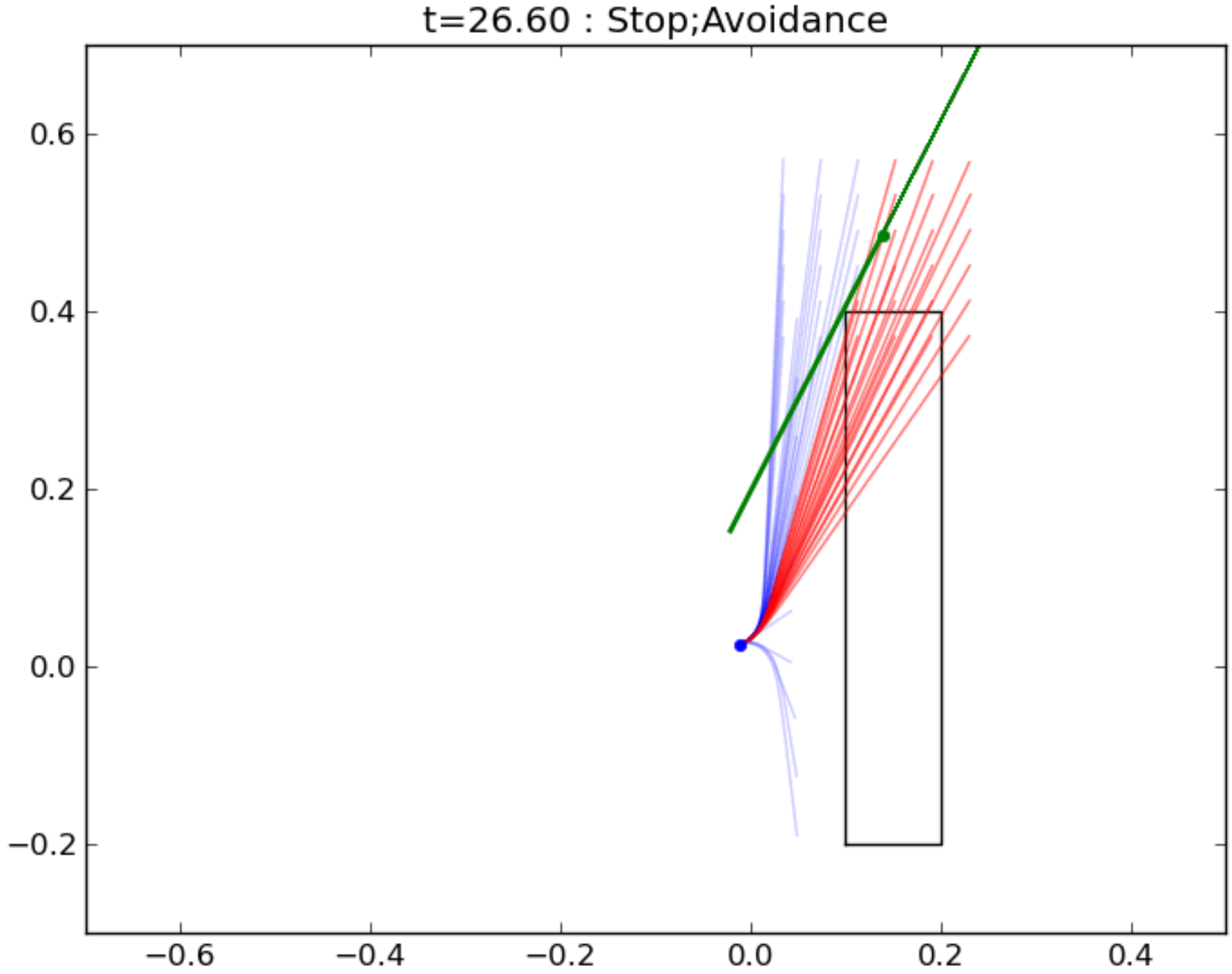


brl
Bristol Robotics Laboratory

Trial 2 – an ethical robot

Bristol Robotics Laboratory

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

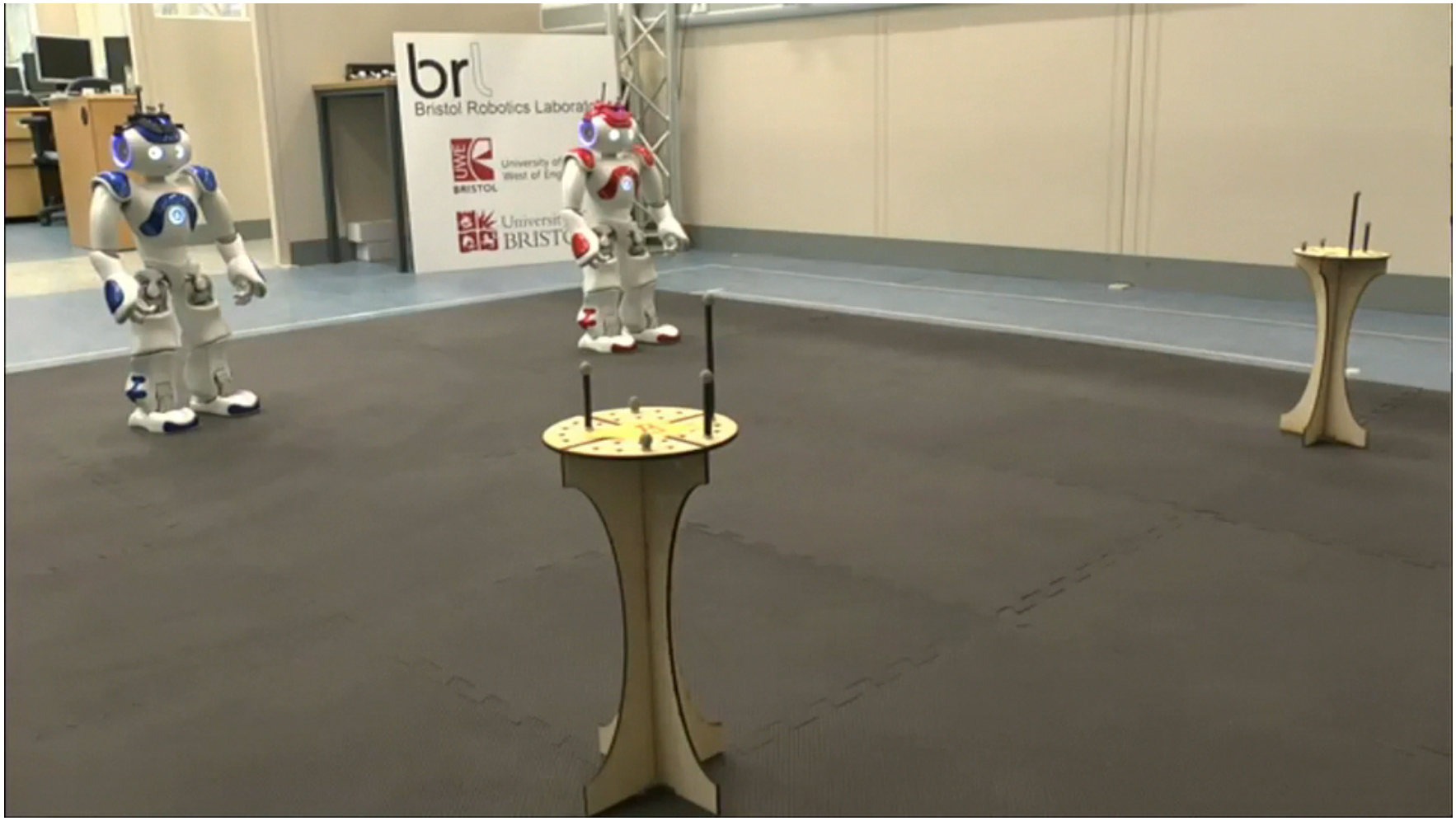

t=26.60 : Stop;Avoidance
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.4
−0.2
0.0
0.2
0.4

Trial 3: the robot's dilemma

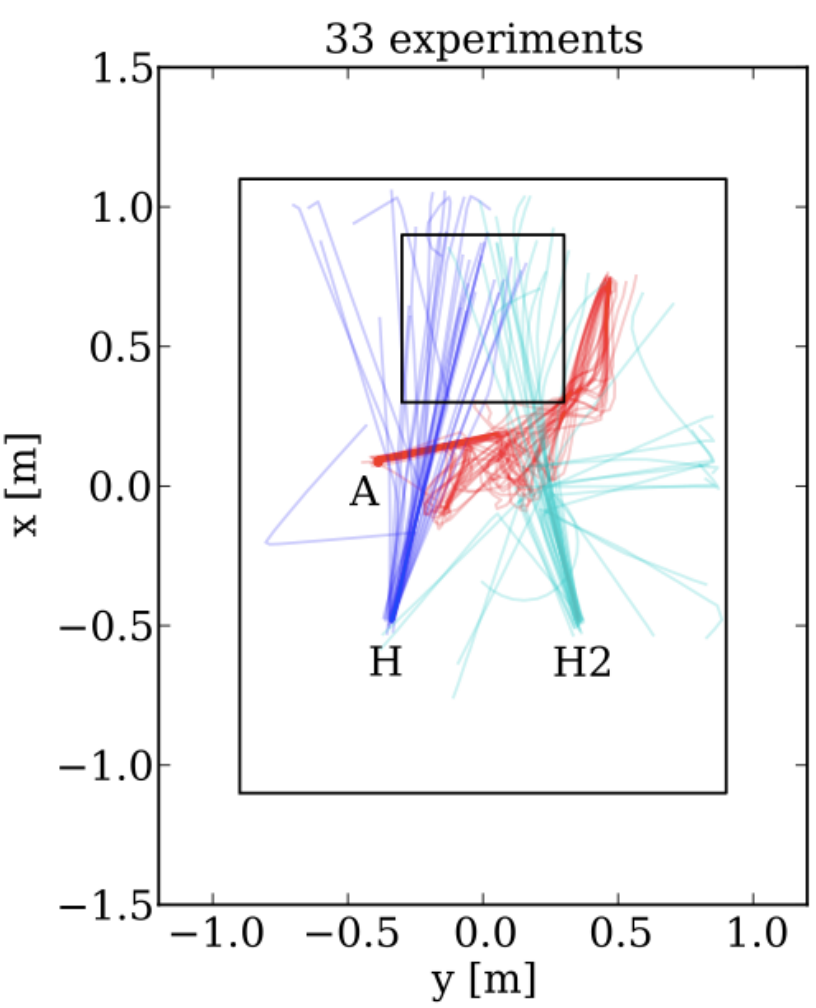
Bristol Robotics Laboratory

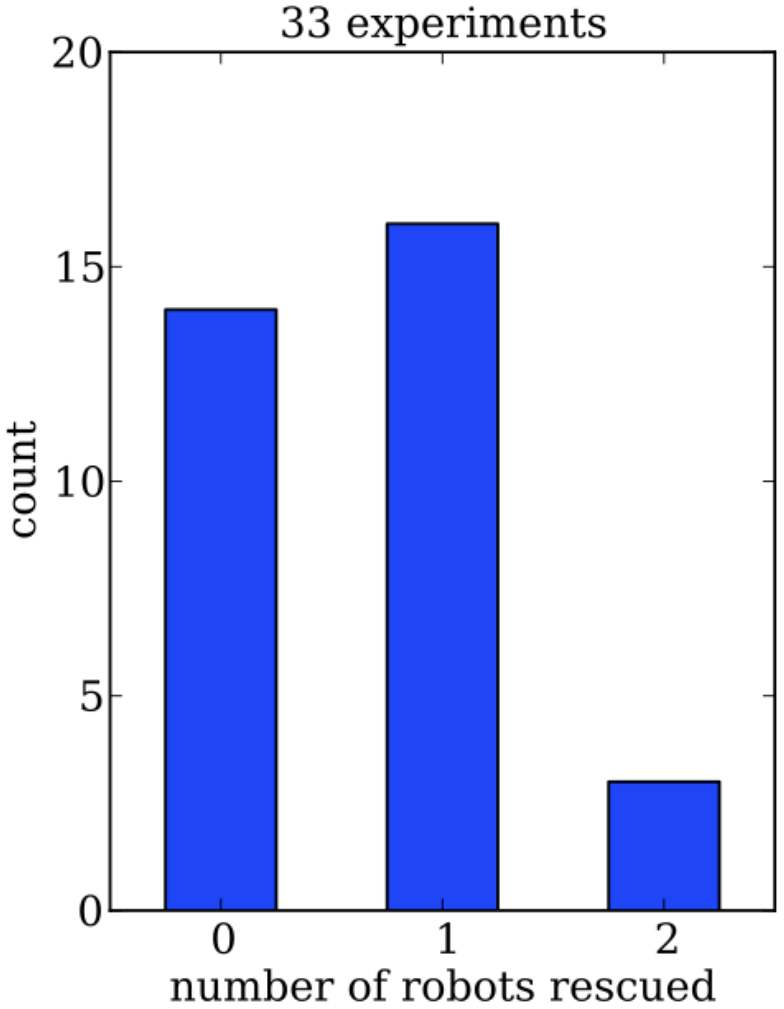
https://drive.google.com/?authuser=0#folders/0BwjY2P_eeOeiQUktZHBBRnozY3M

NAO implementation



Test results: trial 3, an ethical dilemma





In conclusion

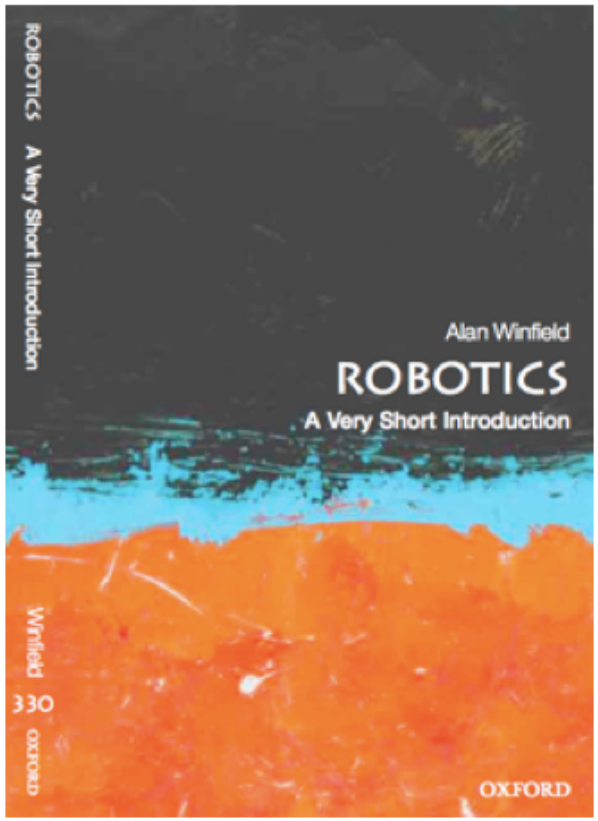
- We must build *safe* cognitive systems
  - able to cope with uncertainties and unpredictable environments…
- Such systems need *situational awareness*
  - Internal models provide a powerful generic architecture which we could all situational imagination
- Self- and other-simulation, in real-time, moves us toward safer (and ethical) systems in unpredictable environments with other dynamical actors

brl
Bristol Robotics Laboratory

Thank you!

- References:
  - 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, LNCS Vol 8717, Springer, 2014.
  - Dennis LA, Fisher M and Winfield AFT (2015), Towards Verifiably Ethical Robot Behaviour, Proceedings of the 1st International Workshop on AI and Ethics, Austin, Texas, 2015.
- 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, Dr Christian Blum and Dr Dieter Vanderelst

EPSRC
Engineering and Physical Sciences Research Council



brl
Bristol Robotics Laboratory