Towards an ethical black box

University of the West of England BRISTOL

UNIVERSITY OF OXFORD

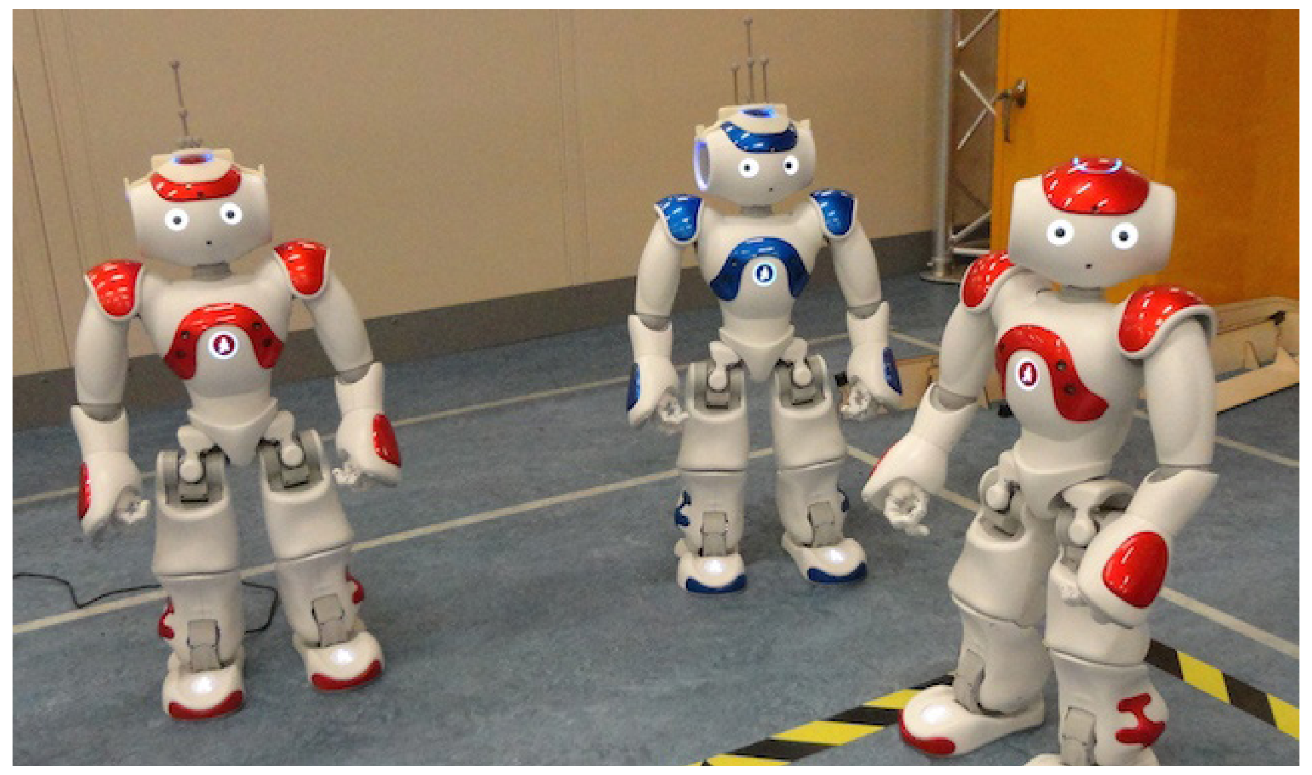
Alan Winfield, Marina Jirotka & Luke Hutton

Recent press articles have reported on accidents in driverless cars. While the accidents were "minor" and apparently not the fault of the car's AI, there is a worrying perception that transparency is lacking in how these events are disclosed. System developers speak reassuringly but we may suspect they have a vested interest in giving events a positive gloss. This raises the crucial question of how the transparency of robot AIs can be guaranteed so as to avoid publics becoming fearful and to ensure that robots gain high levels of trust and acceptance in society.

We propose work to develop ethical black boxes for robots to address these issues head-on. In this project we will anticipate the new technologies and governance structures that will be needed to ensure the transparency and accountability of pervasive, socially embedded and highly autonomous robotic systems.

Research questions

Black boxes - or flight data recorders - have vastly expanded in scope in what flight data they record. Our approach will be to extend black box functions to recover the AI decision-making process as well as environmental factors occurring prior to an incident. The key technical aims of this project will be to develop a proof-of-concept ethical black box (EBB), together with tools to intelligently replay the data collected by the black box, and test the concept with mock robot accident investigations.

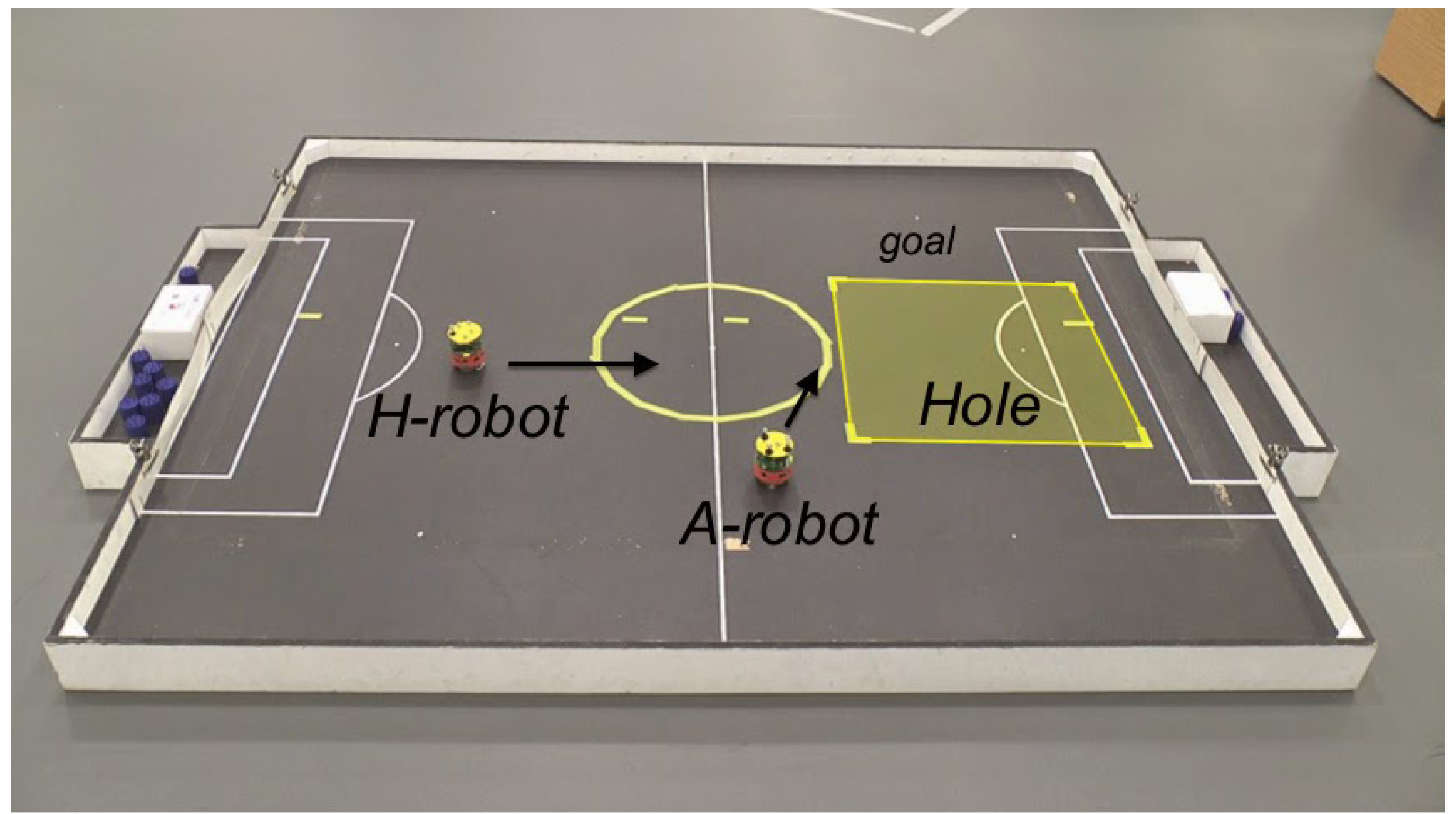


While flight recorders are witnesses to events that often would otherwise remain unwitnessed, the activities of robots are likely to take place in populated spaces where there may be many witnesses, some of whom will record, share and publish details of the event via mobile phones and other devices, creating multiple perspectives on what may have happened. How does interpretation of robot black box data sit alongside the interpretation of evidence from other witnesses, and what is the epistemological status of different types of witnessing?

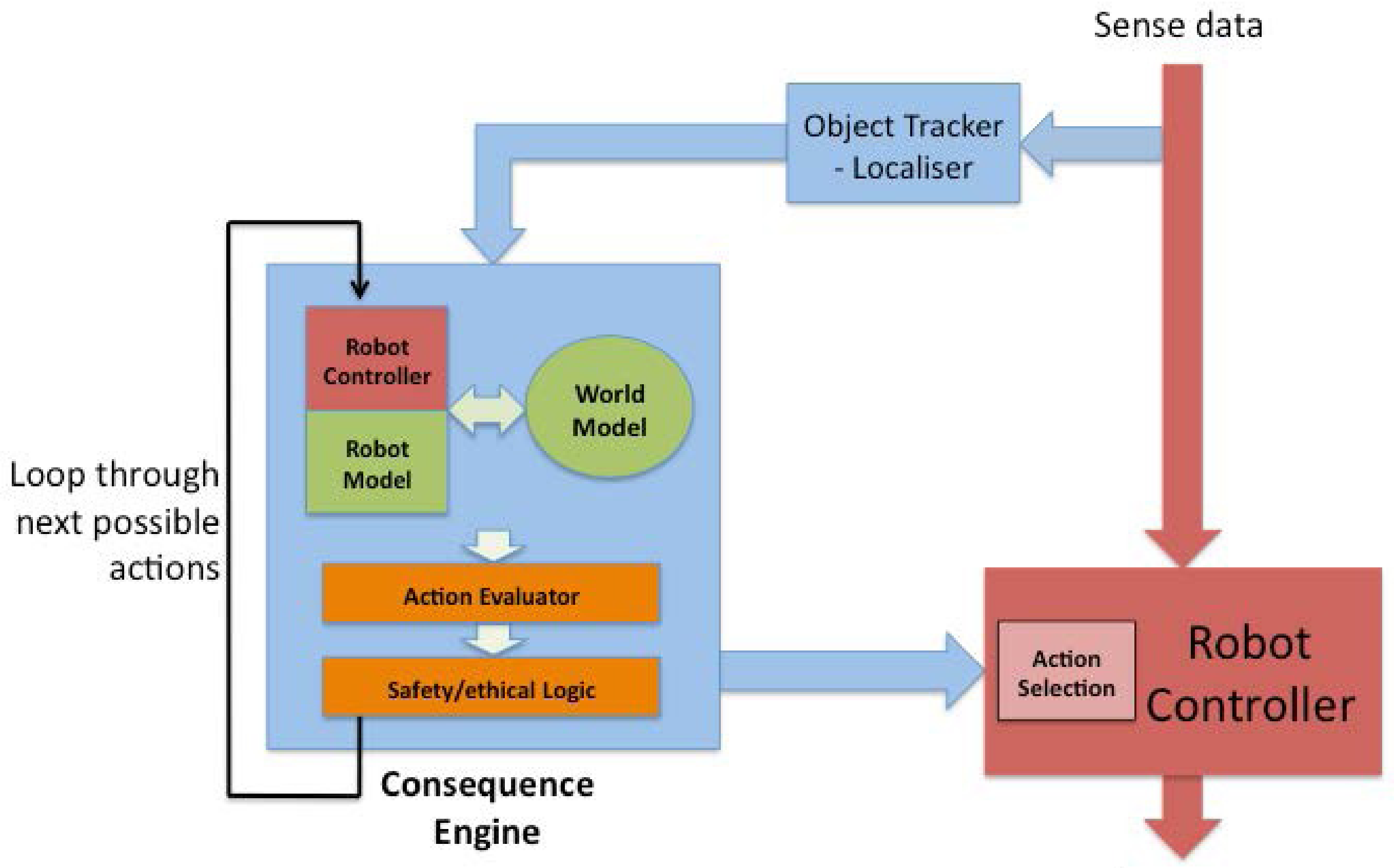
What data is it necessary to store in the EBB to support a post-hoc investigation into the robot's behaviour? Which parts of the reasoning process can be captured, and how can they be presented intelligibly to investigators? What kinds of environmental or contextual data might the EBB collect - such as what the robot 'sees'?

What are the ethical implications and necessary safeguards associated with collecting this data and different strategies for how it might be stored and accessed? What governance structure would be needed for securing data within a black box and for its subsequent handling in any investigation? What might the implications be of a 'glass box' solution where data resides in the cloud or on the manufacturer's servers?

What are the benefits and limitations of a robot incident investigation based on data collected by the EBB? How successfully can such an investigation reach conclusions about why the robot behaved the way it did, and make recommendations? What governance structures are suggested by the staged robot investigations?



Here the part of the human is played by another robot, the H-robot, which is heading toward the (virtual) hole in the experimental arena, while the A-robot has both an internal model (simulation), a consequence evaluator and a safety/ethical logic layer which allows it to select its own actions based on rules hard-coded into that logic.



Internal mode based architecture of robot with Consequence Engine. Robot control data flows are shown in red, while internal model data flows are in blue.

Ethical robots

If you see someone walking toward a hole in the ground, perhaps distracted by their phone and not looking where they are going, you are likely to intervene. The reason you are able to do so, and hopefully prevent that person falling into the hole, is not just because you are a good citizen; it is because you have the cognitive machinery to predict the likely consequences of her actions (if you do nothing) and your actions (and interaction with her) if you intervene.

We have developed a robot with an internal model – a simulation of itself, its environment and other dynamic actors inside itself – and experimentally tested this scenario.

We call this internal model a **consequence engine.**

Consequence engine

The consequence engine runs in real-time. Every 0.5 seconds it simulates 30 next possible actions of the A-robot (i.e. different directions of movement) for 10 seconds into the future. For each of those 30 next possible actions the H-robot is also simulated, by assuming it will continue to move in the direction and with the same speed it is currently moving. The simulated outcomes for both the A and H robots are given numerical values according to the level of danger in those outcomes, and safety/ethical logic is applied to choose the A-robot's next action.

Experimental trials have demonstrated that the A-robot does indeed intervene to prevent the H-robot from falling into the hole, before then continuing toward its own goal position while maintaining its own safety.

Introducing a second H-robot also heading toward the hole presented the A-robot with a balanced ethical dilemma. Here the A-robot was unable to prevent either of the two H-robots from falling into the hole in almost half of the experimental runs. In fact the A-robot demonstrated unexpected "dithering" behaviour – unable to decisively choose between H-robots to rescue it wasted time and was unable to rescue either. Future real-world ethical robots are likely to face difficult, if not impossible, ethical dilemmas that are likely to be very challenging to analyse post-hoc.