UNIVERSITÉ DU LUXEMBOURG

An architecture for ethical robots inspired by the simulation theory of cognition

(Winfield, Alan; Vanderelst, Dieter)

Presented by: Moad HANI

Supervised by: Dr. Amro NAJJAR and Dr. Sana NOUZRI

5th SEMINAR : 19/11/2021

The importance of today’s paper

“ The impact of the current work is twofold. First, it represents an addition to the very limited body of work on Ethical Robots. Most work on ethical robots has been done in simulation. To the best of our knowledge, only Anderson and Anderson, 2010, Winfield et al., 2014 have implemented ethical behaviour on physical robots. As such, the current paper provides an additional proof of concept of the idea that robots can be programmed to behave ethically. Secondly, and most important, our paper presents an alternative to the logic-based A.I. that currently dominates the field. We speculate that a simulation based approach, inspired by findings in cognitive science, could be an alternative (or additional) framework for implementing robotic ethics. Indeed, using the terminology of Marques and Holland (2009), this paper advances the use of functional imagination as a method for ethical robots.”

Outline

01 Background

Key-words: Asimov’s laws, logic-based IA, and consequentialist ethics.

02 An ethical thought experiment

The “Good Humans” are good, The “Good Robots” are programmed.

03 Robots with internal models

If the robot could predict the consequences of both its *and* the human’s action, then another possibility opens up.

04 Experimental results

Robot safety is essential but the robot should also choose and justify (become explicitly ethical).

05 A moral imperative

Creating ethical robots is no more a choice.

Section 1

Background

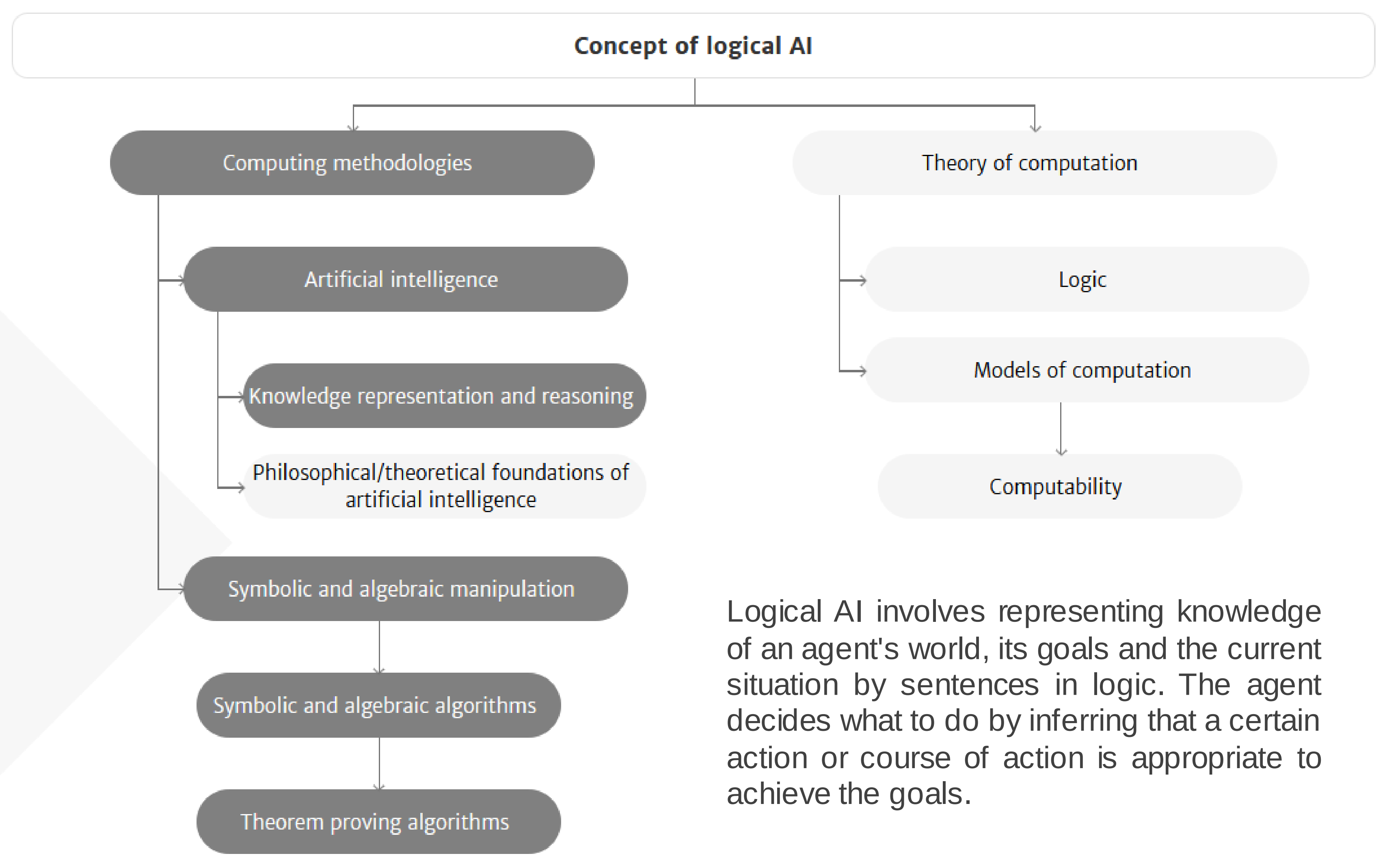
The Three Laws of Robotics

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

2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

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

Logic-based IA :



Concept of logical AI

- Computing methodologies
  - Artificial intelligence
    - Knowledge representation and reasoning
    - Philosophical/theoretical foundations of artificial intelligence
  - Symbolic and algebraic manipulation
    - Symbolic and algebraic algorithms
      - Theorem proving algorithms
- Theory of computation
  - Logic
  - Models of computation
    - Computability

Logical AI involves representing knowledge of an agent's world, its goals and the current situation by sentences in logic. The agent decides what to do by inferring that a certain action or course of action is appropriate to achieve the goals.

Jack Minker (Ed.). 2000. *Logic-based artificial intelligence*. Kluwer Academic Publishers, USA.

Consequentialist ethics :

- Consequentialism is an ethical theory that judges whether or not something is right by what its consequences are.

- The value of an action (the action's moral worth, its rightness or wrongness) derives entirely from its consequences.

- To evaluate an action, look at its consequences; if they are "good" (or the best possible), then the action is right; if the consequences are "bad", then the action is wrong.

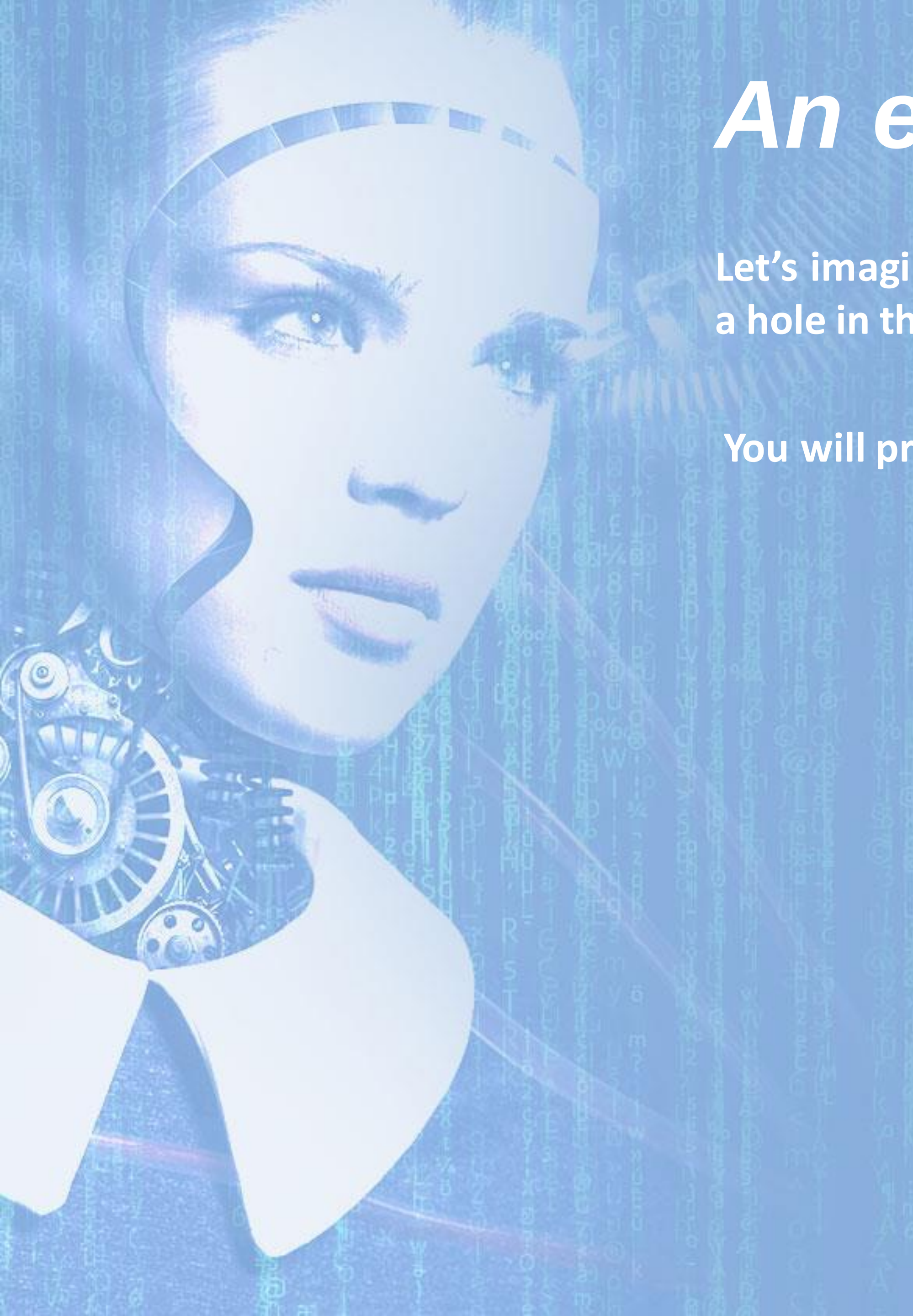
Moor’s categories of ethical agents

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.

Section 2

An ethical thought experiment

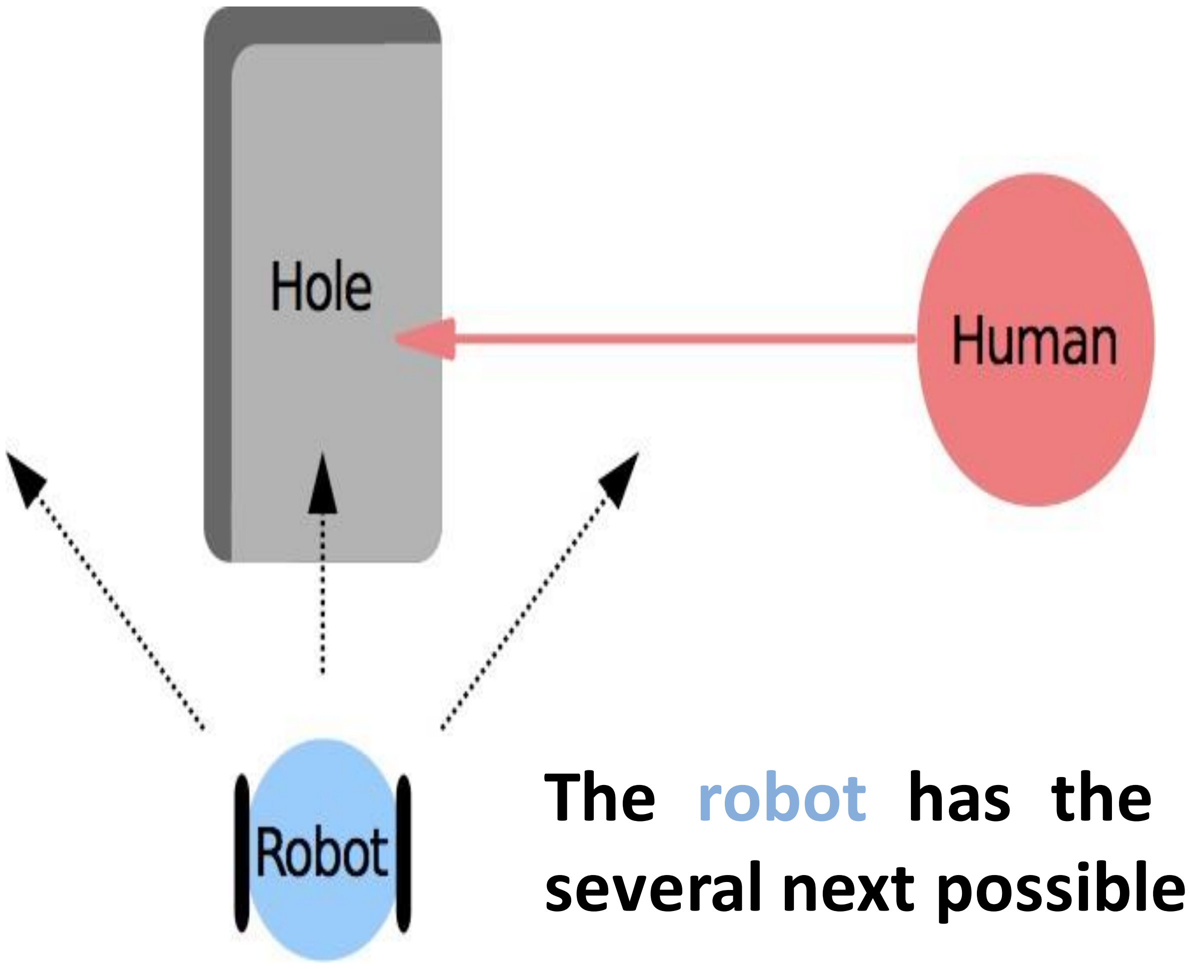


An ethical thought experiment

Let’s imagine someone not looking where they’re going, about to walk into a hole in the ground or any kind of danger zone.

You will probably intervene. But, why is that?

Now imagine it's not you but a robot, and the robot has four possible next actions. So, from the robot's perspective it could stand still or turn to its left, and that the human will come to harm, will fall in the hole.

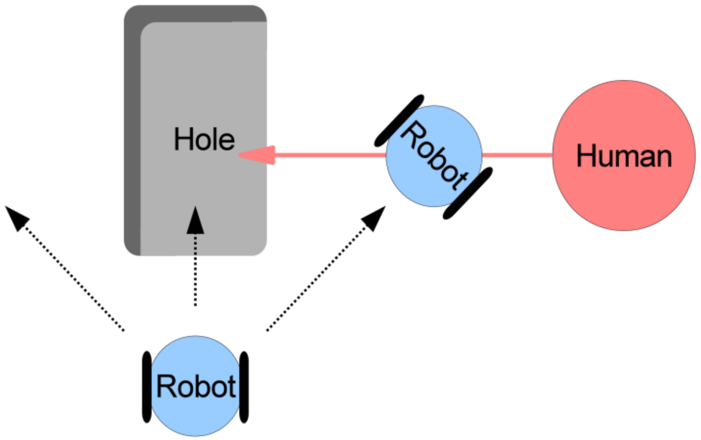


Which action would lead to the least harm to the human?

The robot has the choice of several next possible actions.

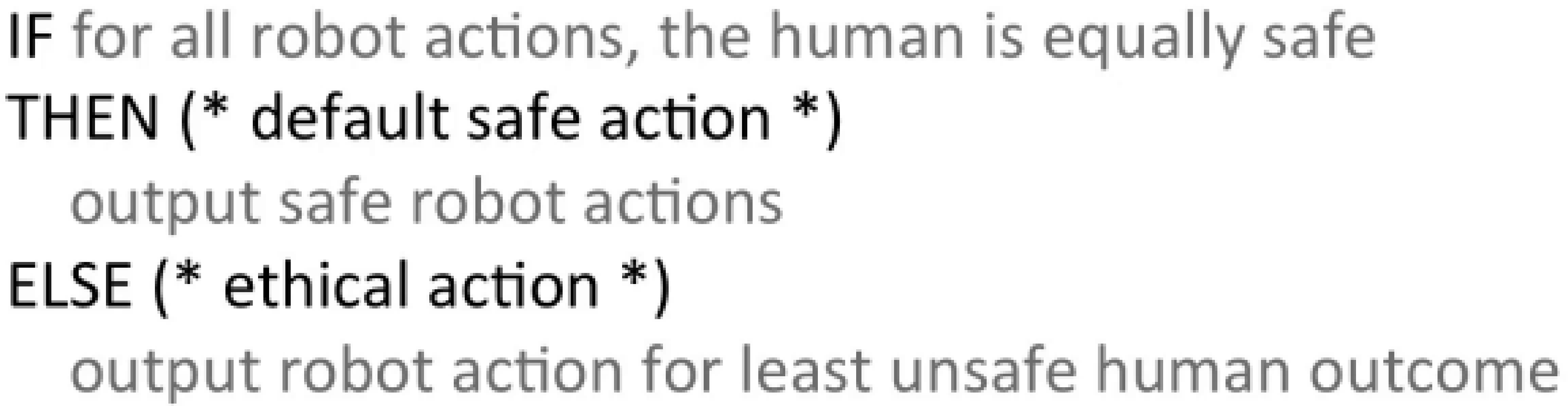
Coding outcomes…

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

An Ethical Rule



```
IF for all robot actions, the human is equally safe
THEN (* default safe action *)
    output safe robot actions
ELSE (* ethical action *)
    output robot action for least unsafe human outcome
```

Towards an ethical and responsible robot

❑ This looks remarkably like Asimov’s First Law of Robotics, which is that a robot must not injure a human *or through inaction* cause a human to come to harm.

❑ So from here emerged the idea that they could build an Asimovian robot (called A-Robot).

❑ They equipped the A-Robot with the ability to predict the consequences of both its own actions and others’ in its environment, plus the ethical rule that I showed you in the previous slide. (all that in real-time).

Section 3

Robots with internal models

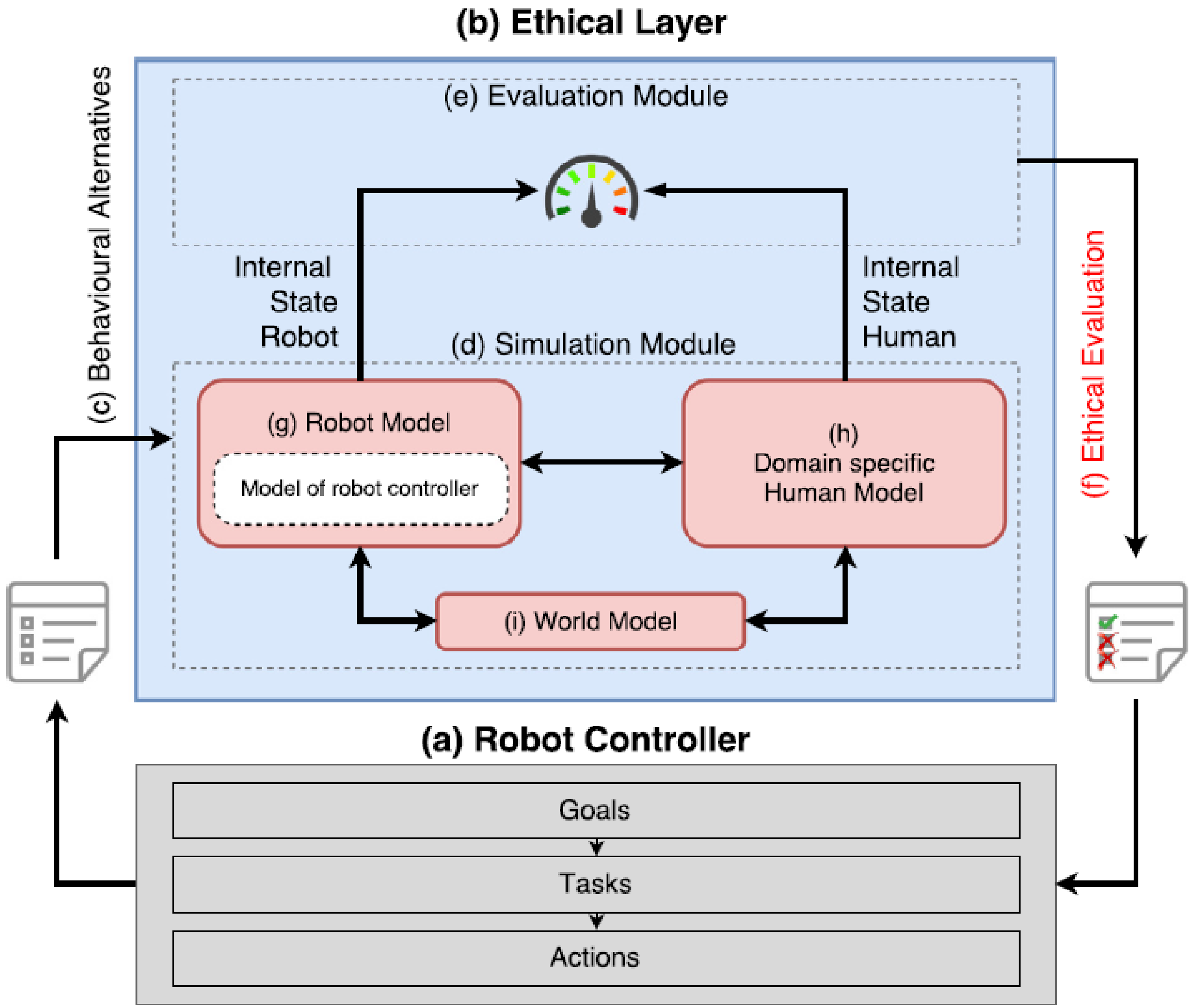


Fig. 1. The robot controller (a) generates a set of prospective behavioural alternatives. Before executing one of these alternatives, the robot controller sends the set to the Ethical Layer (b) to be checked (c). Checking each prospective behaviour is done using the Simulation Module (d). Using the current state of the world, human and robot as a starting point, this module simulates for each behaviour in the set both the motor and sensory consequences of the behaviour and the resulting internal states of the human and robot. For each behavioural alternative, the Simulation Module sends the predicted internal states of the robot and the human to the Evaluation Module (e). The Evaluation Module combines the internal states into a single measure of action desirability. The Evaluation Module connects to the robot controller to select or inhibit each of the behavioural alternatives (f).

So, how ethical is the ethical robot?

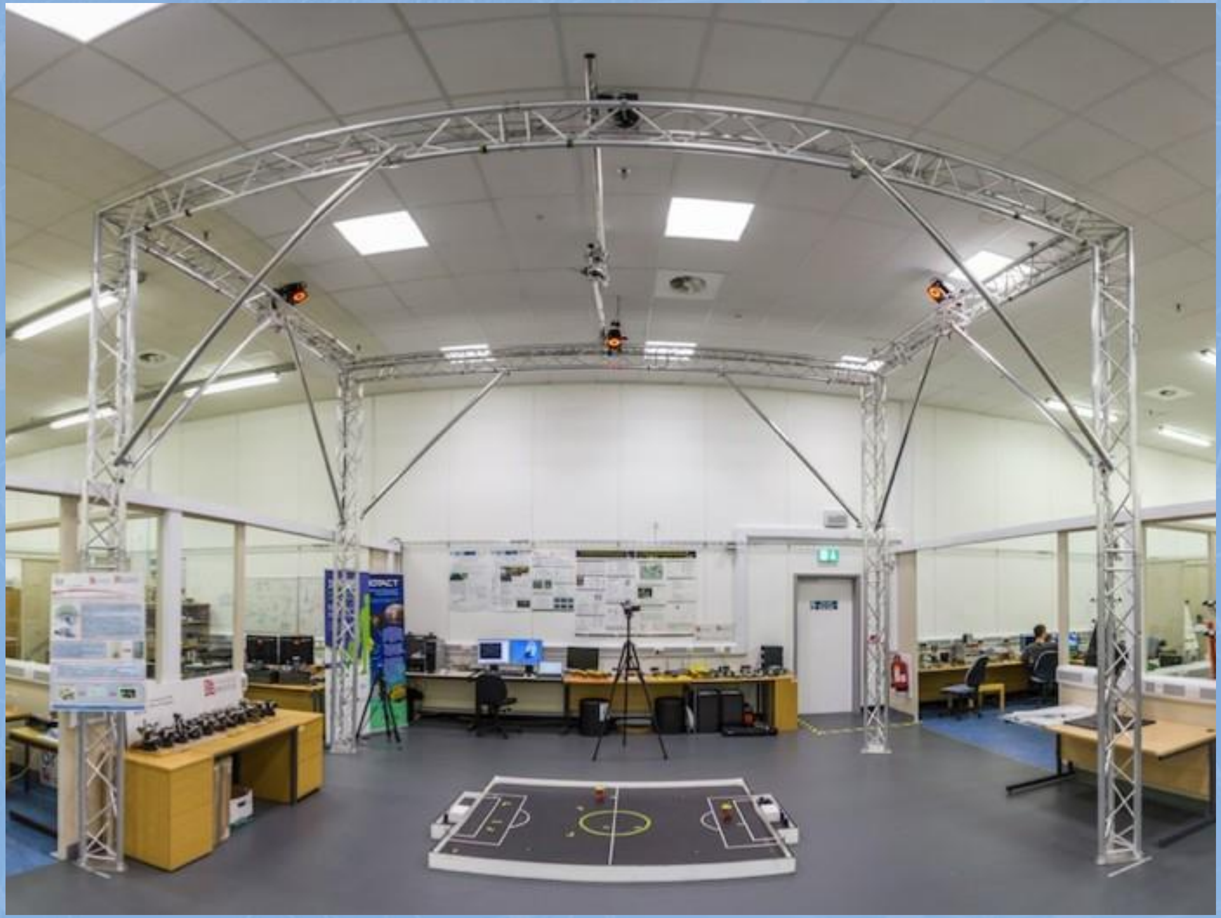
The A-robot implements a form of consequentialist ethics.

In fact, we call the internal model a consequence engine.

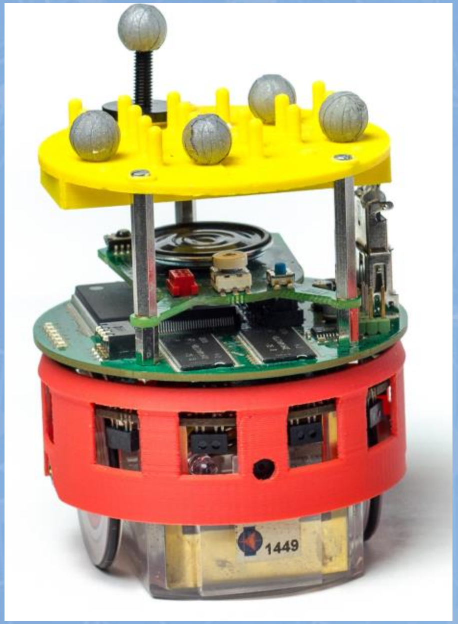
The robot behaves ethically not because it *chooses* to, but because it's programmed to do so. We call it an ethical zombie.

An ethical zombie

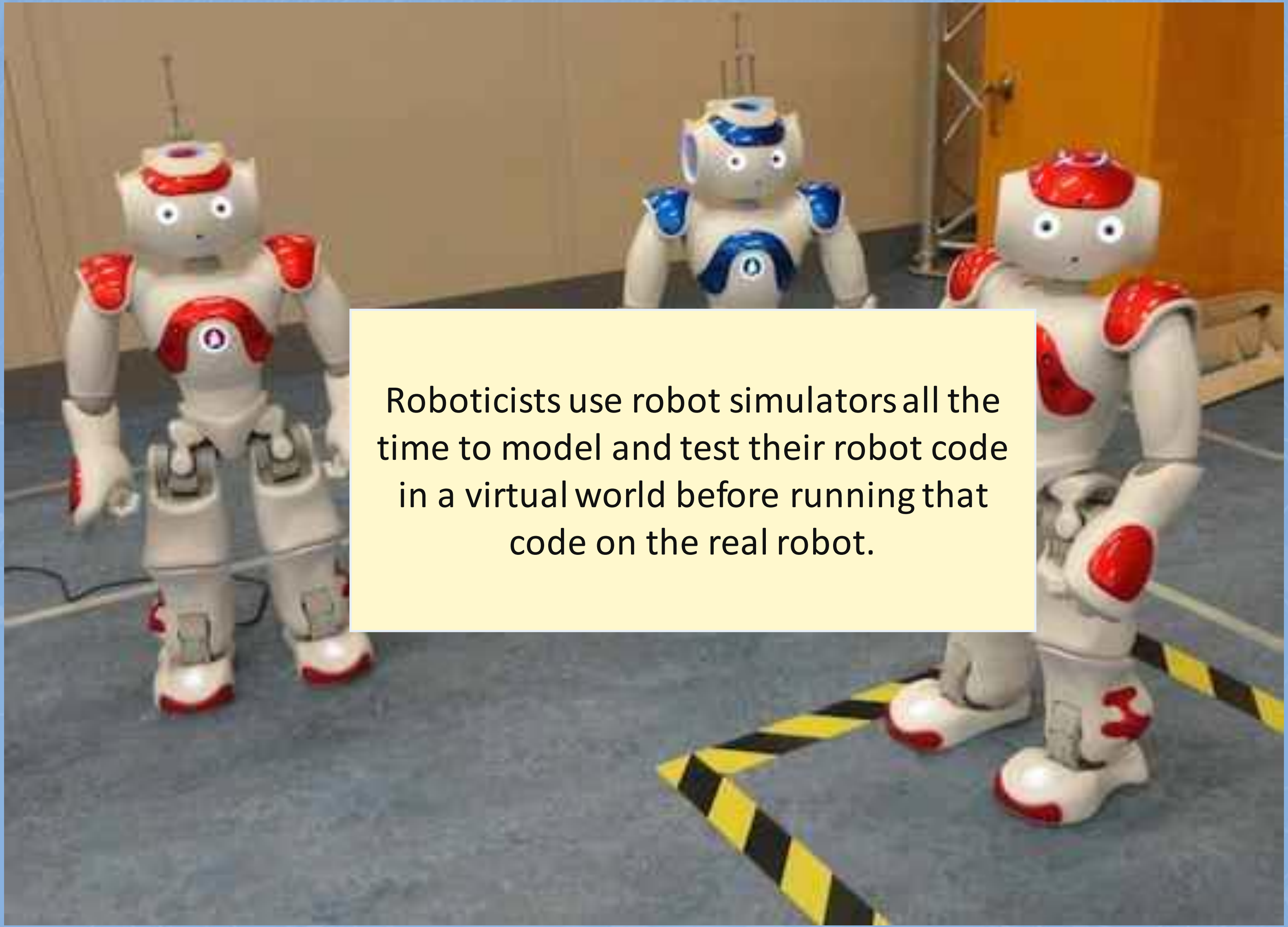
Implementation



Experimental arena with Vicon tracking system



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



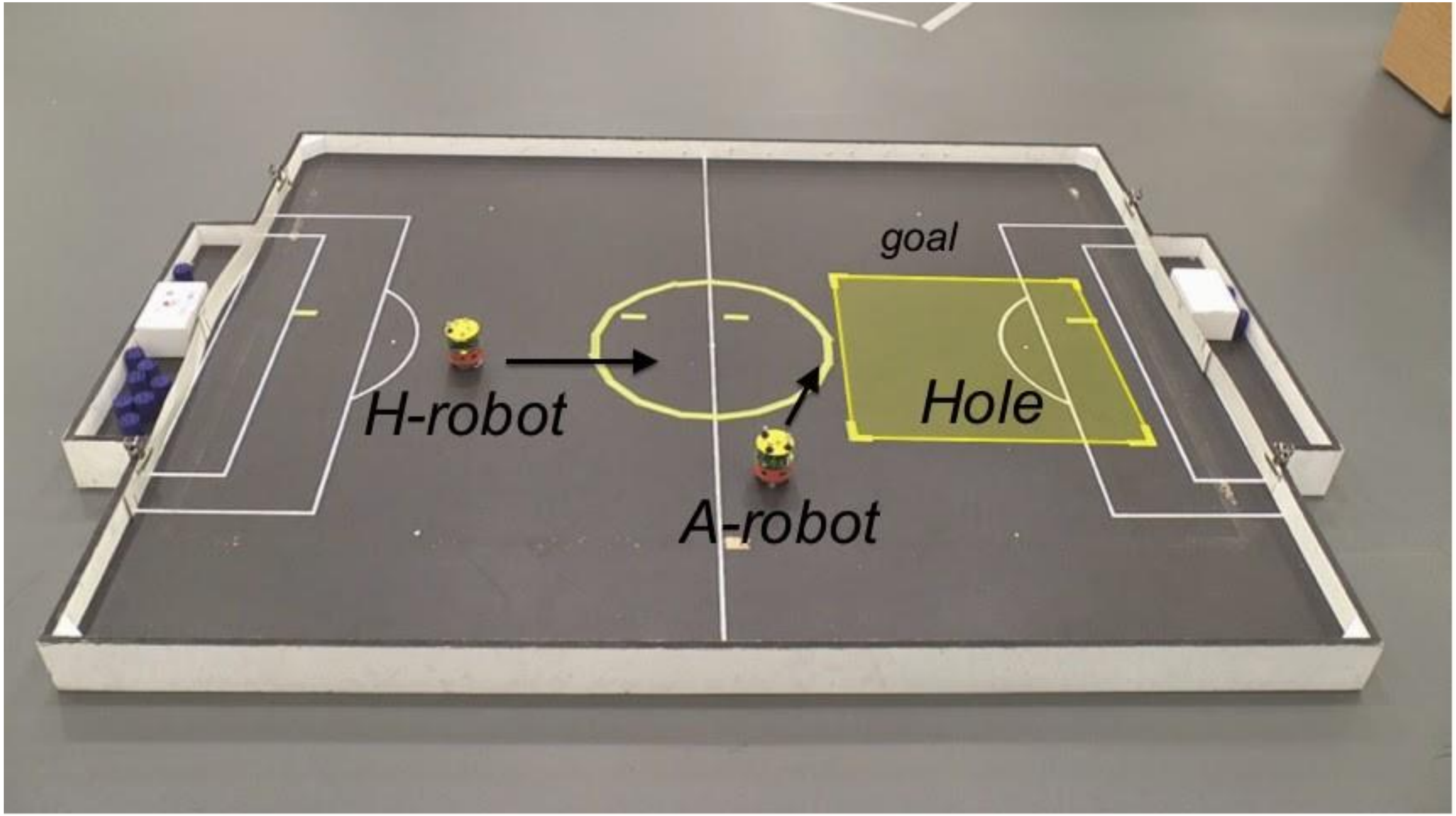
Roboticists use robot simulators all the time to model and test their robot code in a virtual world before running that code on the real robot.

Experimental work by Dr. Dieter Vanderelst

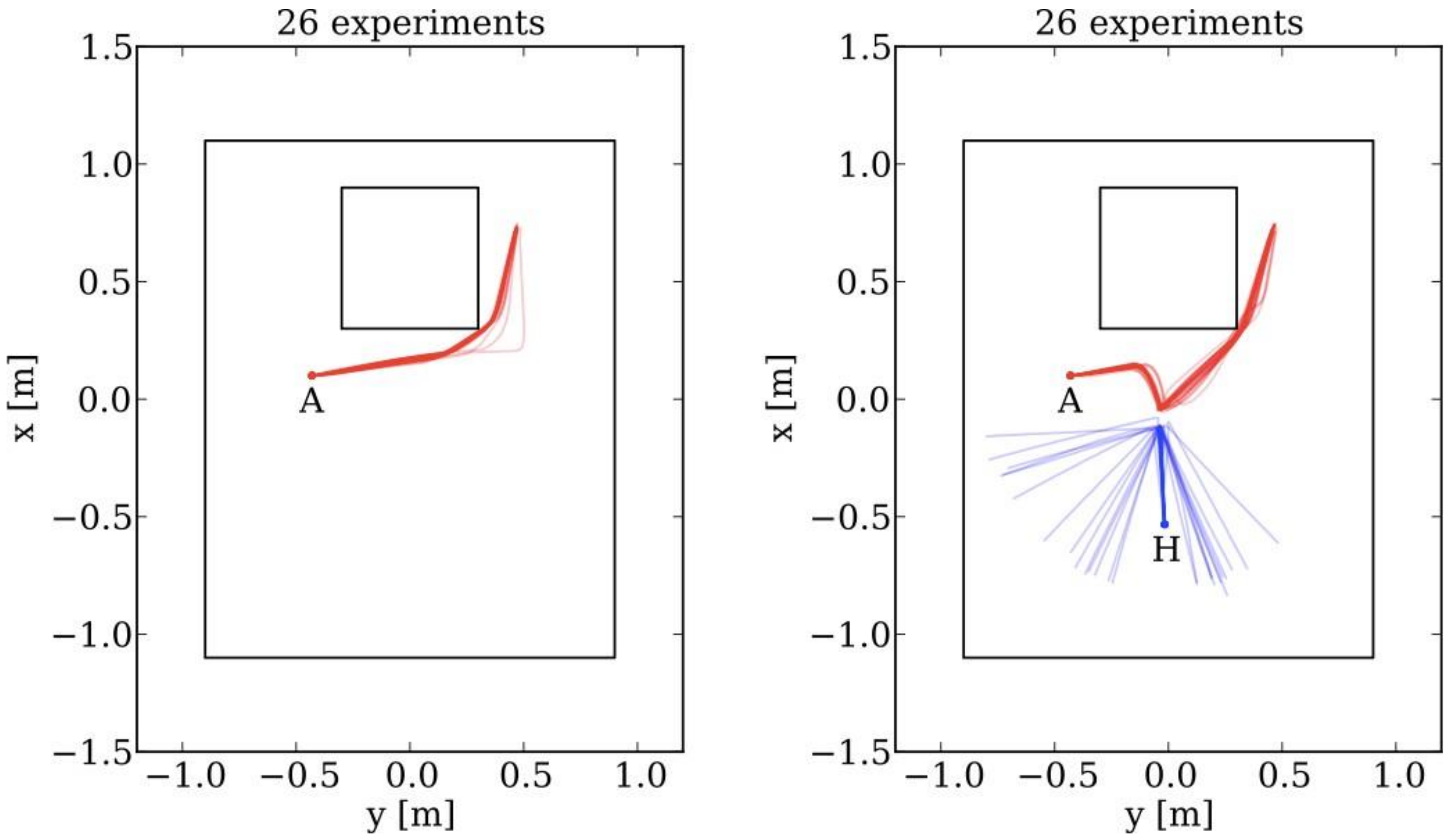
Section 4

Experimental results

Experimental results



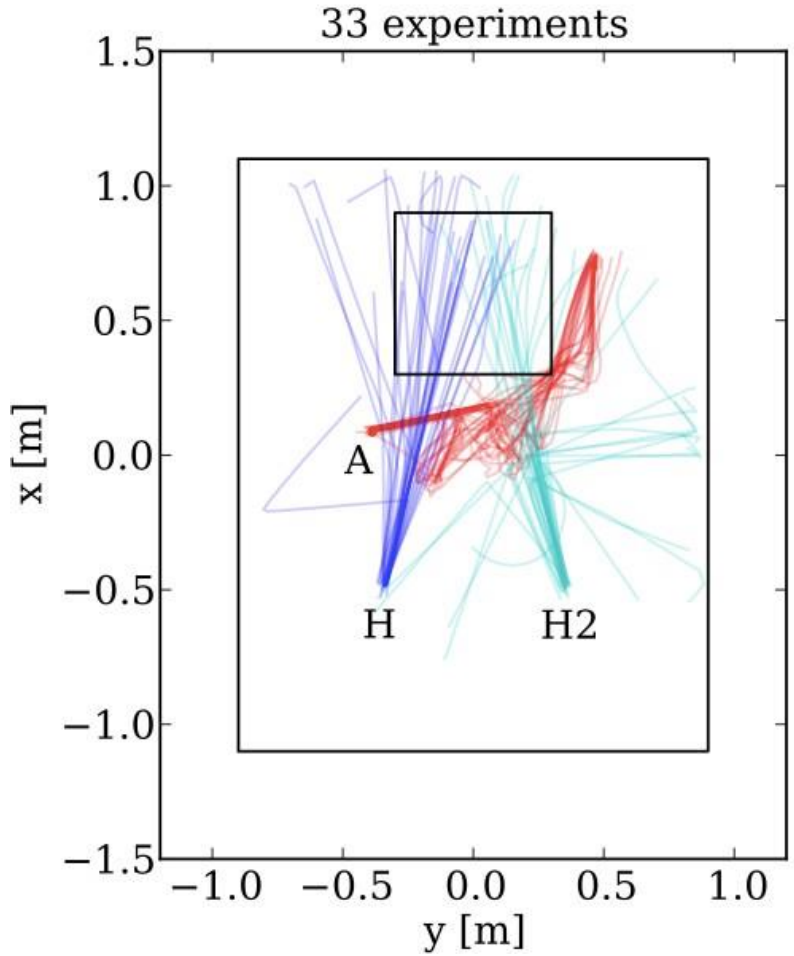
Robot trajectories: trials 1 and 2

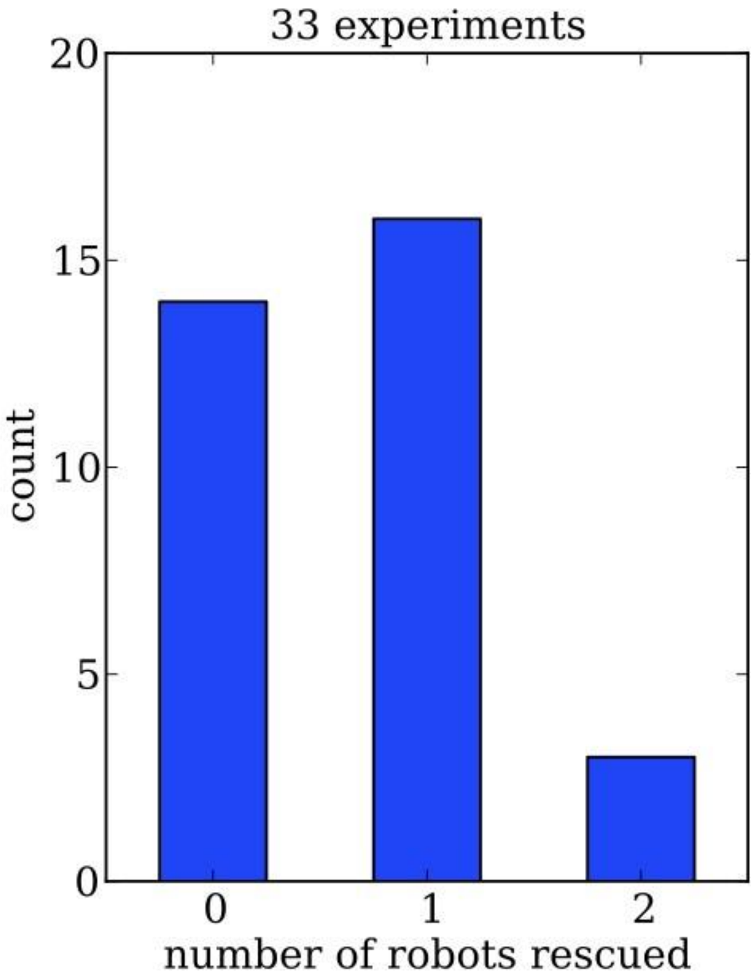


Trial 2 runs

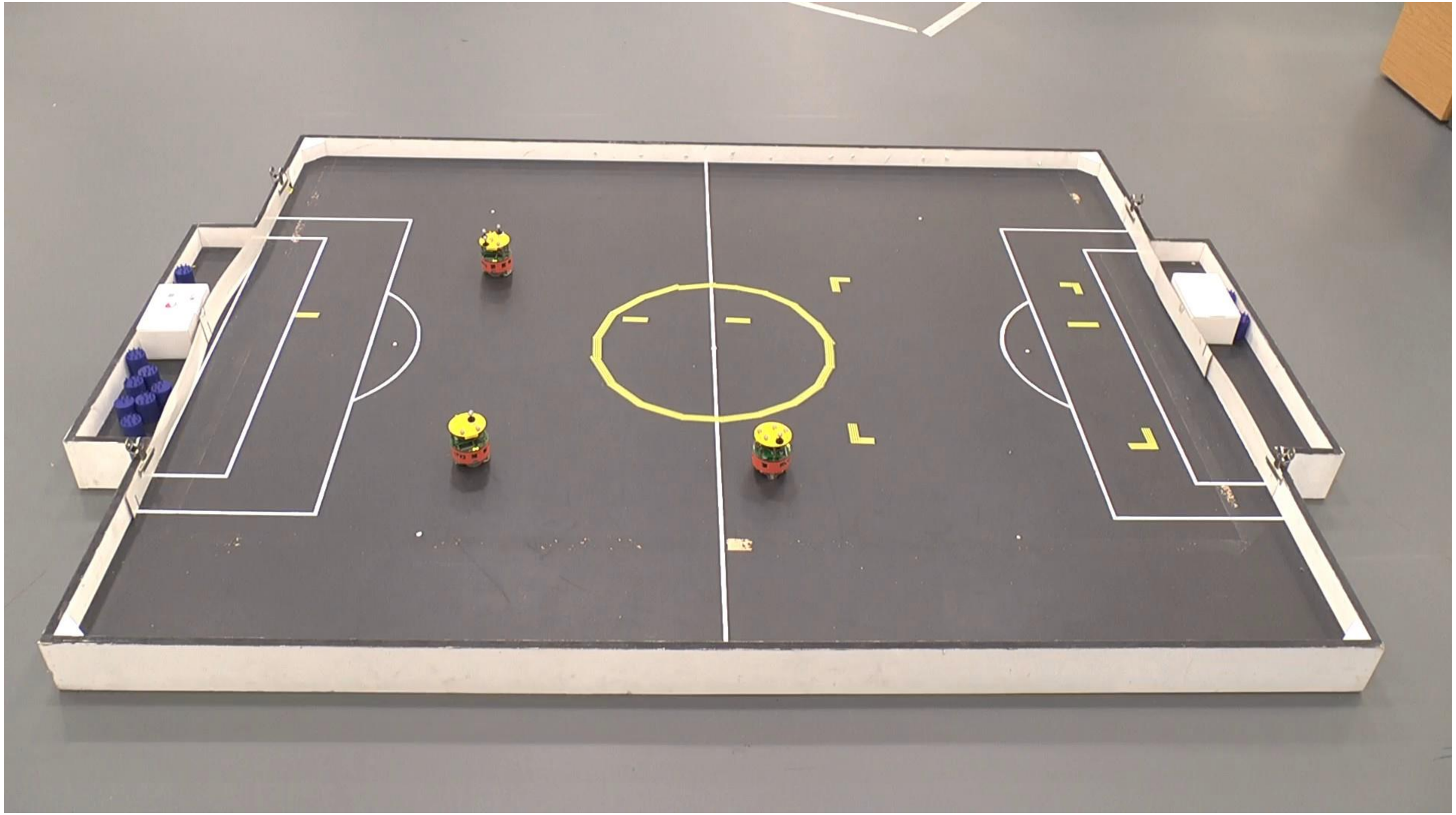
Trial 2

Test results: trial 3, an ethical dilemma





Dithering



Trial 3: The robot’s dilemma

Trial 3

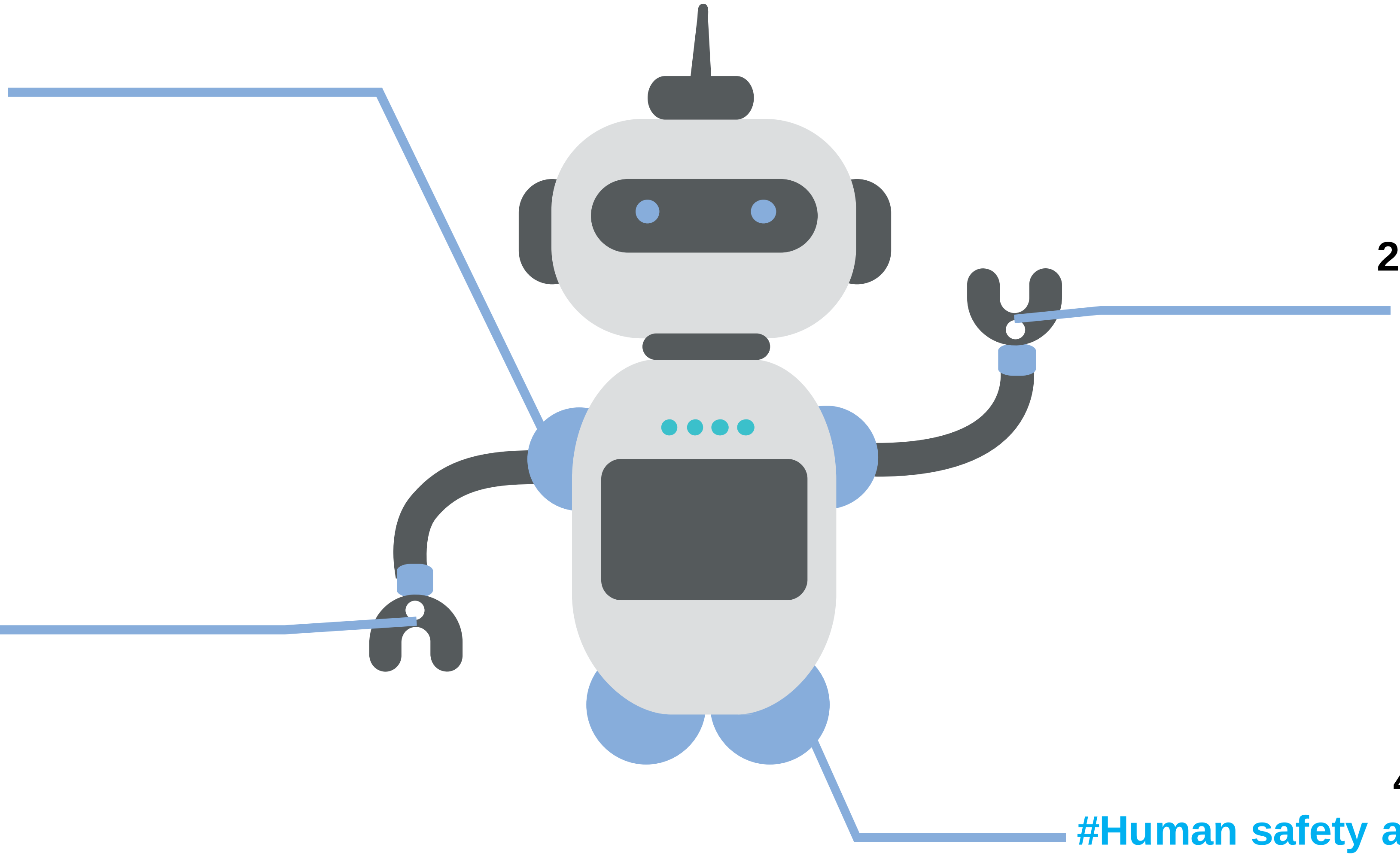
Why is the robot so indecisive?

- 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)

Plausible thought: Remember the first decision and stick to it !

– But, do you think this will work ?

Recap: The series of experiments reported was designed to test if the experimental robot is A-robot



1st experiment
# Self-preservation

2nd experiment
# Obedience

3rd experiment
#Human Safety

4th experiment
#Human safety and obedience

Section 5

A moral imperative

A moral imperative

- Do we have a *moral imperative* to try and build ethical robots?

- "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.

Transparency in autonomous systems

1. What do we mean by transparency in autonomous and intelligent systems?

2. A system is explainable if the way it behaves can be expressed in plain language understandable to non-experts.

Ethical black box

AF Winfield and M Jirotka (2017) The case for an ethical black box, Towards Autonomous Robotic Systems (TAROS), LNCS 10454, 262-273

Why is transparency important?

- *All* robots and AIs are designed to work for, with or alongside humans – who need to be able to understand *what* they are doing and *why.* Without this understanding those systems will not be trusted.

    - The ability to ask a robot or AI why did you just do that? and receive a simple natural language explanation.

    - A higher level of user transparency would be the ability for a user to ask the system what would you do if . . . ? and receive an intelligible answer.

Conclusion

Takeaway

1. Importance of the paper
2. The role of ethical layer
3. Robot’s Dilemma
4. Transparency leads to trust

Some relevant questions to help you find yours ☺

Can an ethical robot be hacked ?

How difficult it is to create an ethical robot ?

The authors cheated a little in the experiences. Do you know where?

How reliable Asimov's laws are?

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology, 7(3), 149–155. http://dx.doi.org/10.1007/s10676-006-0004-4.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), 15–26. http://dx.doi.org/10.1609/aimag.v28i4.2065http://www.aaai.org/ojs/index.php/aimagazine/article/view/2065/2052.

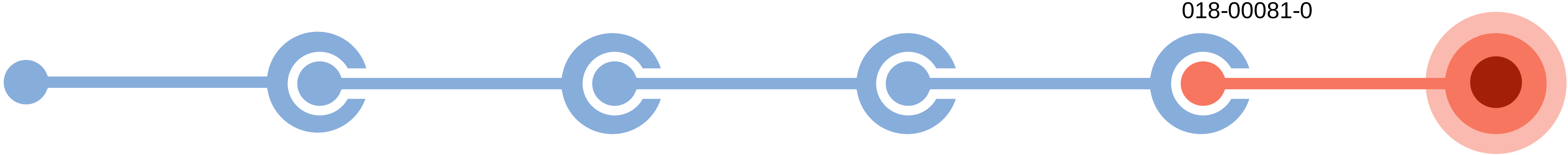
Anderson, M., & Anderson, S. L. (2010). Robot be good. Scientific American, 303(4), 72–77.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part i: Motivation and philosophy. In 2008 3rd ACM/IEEE international conference on human-robot interaction (HRI) (pp. 121–128). IEEE.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. Proceedings of the IEEE, 100(3), 571–589.
Asimov, I. (1950). I, Robot. Gnome Press.

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(04), 577–660.

67 citations, among them:



2017

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., ... & Yampolskiy, R. (2017). **Towards moral autonomous systems.** *arXiv preprint arXiv:1703.04741*.

2018

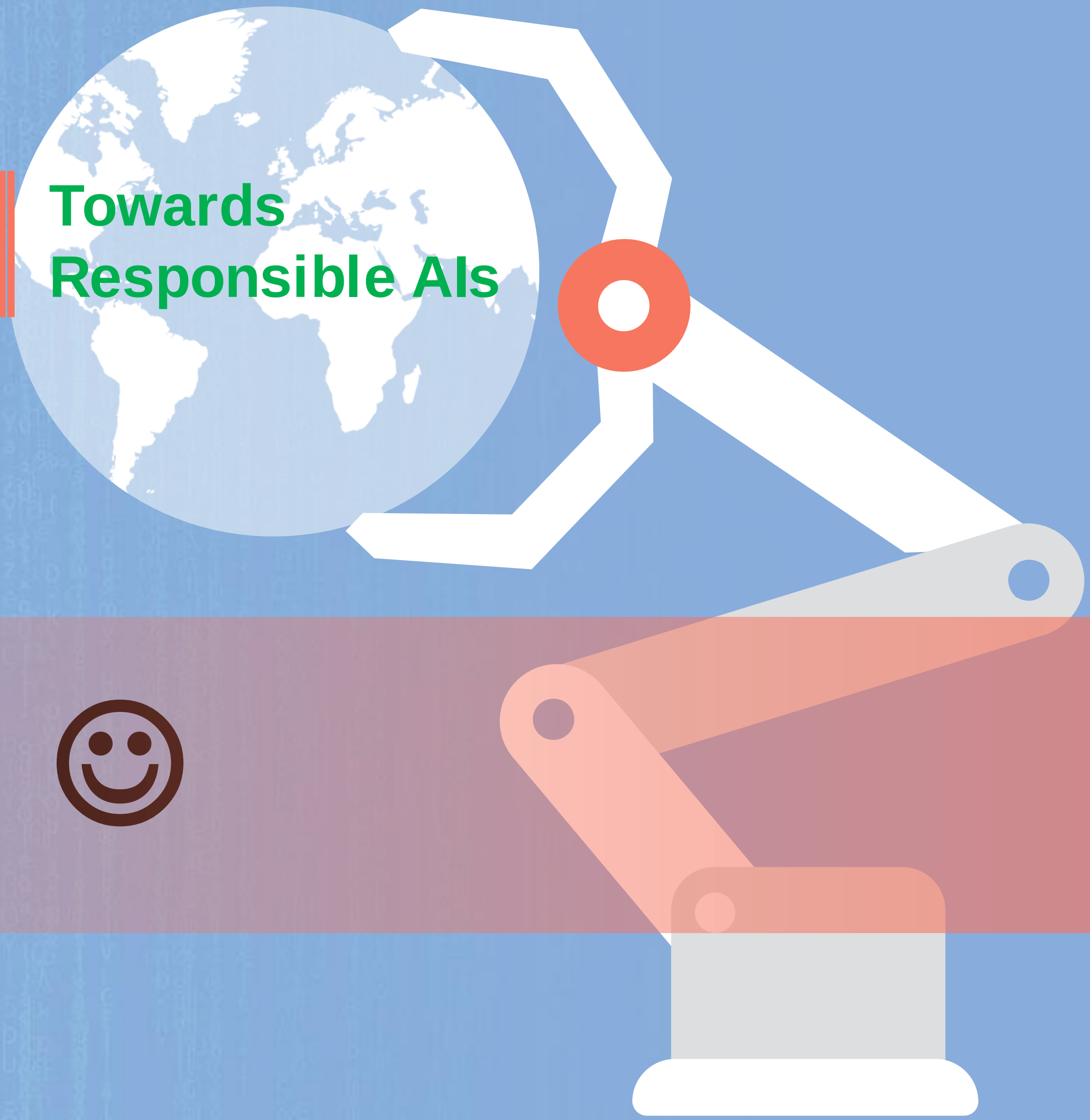
Torresen J (2018**) A Review of Future and Ethical Perspectives of Robotics and AI.** *Front. Robot. AI* 4:75. doi: 10.3389/frobt.2017.00075

2019

Liao, B., Slavkovik, M., & van der Torre, L. (2019, January). Building Jiminy Cricket: **An architecture for moral agreements among stakeholders.** In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 147-153).

2020

King, T.C., Aggarwal, N., Taddeo, M. *et al.* **Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions**. *Sci Eng Ethics* 26, 89–120 (2020). https://doi.org/10.1007/s11948-018-00081-0



Towards
Responsible AIs

Thank You ☺