Transparent Minds: A principled challenge to build trustless AI for social robot applications

Rob Wortham

Department of Computer Science
University of Bath

@RobWortham

AmonI Research Group

UNIVERSITY OF BATH

Joanna Bryson

(Currently Princeton)

Swen Gaudl (Falmouth)

- Game AI
- Reactive Planning
- Genetic Algorithms

Andreas Theodorou

- AI Transparency
- Public Goods

Yifei Wang

(Georgia Tech)

- Bio Evo Models
- GRN

Paul Rauwolf (Oxford)

- Modelling
- Human Biases
- Self Deception

Rob Wortham

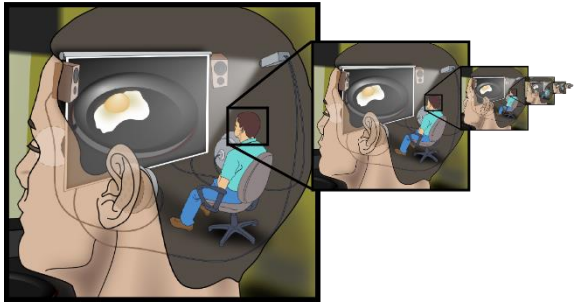
- AI Transparency
- Robots
- Ethics

Today's Talk:

1. Ethical **Principles** for **Designers** of Robots

2. An Experiment Investigating **Robot Transparency** (IJCAI-16 July).

3. Building ***Transparent Minds***



EPSRC Principles of Robotics[1]:

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

1. https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

The Principles are

for this **guy**

Not for this **thing**

Principle Four:

Robots are manufactured artefacts.

- They should not be designed in a **deceptive** way to exploit **vulnerable** users;
- instead their machine nature should be **transparent**.

But why does this matter?

It seems obvious. So what? Why should we care?

Mueller's Transparent Computers[2]…

[Cashpoint machines / word processors / mobile phones / library systems / banking web sites / social networking /…]

1. Promote understanding
2. Educational
3. Easier to fix problems
4. Improves Customer Satisfaction (?)
5. Builds Trust (?? **Confidence**)

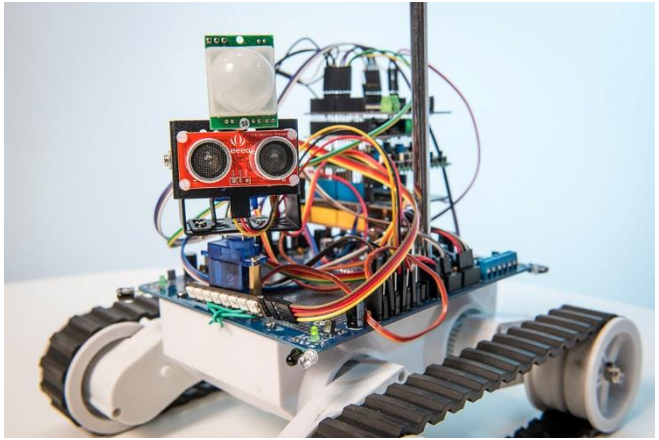
Poor Transparency -> systems that are **difficult** / **frustrating** to use

2. Erik T. Mueller, Transparent Computers: Designing Understandable Intelligent Systems, 2016

Do We *Need* Transparent Robots?

Video - 5 minutes long

- **QUESTION: What do you think the robot is trying to do?**



Experiments

- **QUESTION: What do you think the robot is trying to do?**

- “Trying to create a 3d map of the area? At one stage I thought it might be going to throw something into the bucket once it had mapped out but couldn't quite tell if it had anything to throw.”

- “aiming for the black spot in the picture.”

- “is it trying to identify where the abstract picture is and how to show the complete picture?”

Answers from STEM Graduates - March 2016

EPSRC Principles ……

- They should not be designed in a **deceptive** way to exploit **vulnerable** users;

Humans are not equipped by genetic or cultural evolution to deal with machine agency[3] – we have No Theory of Mind for Robots.

So we make stuff up!

We are all vulnerable users

3. Bryson, J. J., 2012. Patiency is not a virtue: suggestions for co-constructing an ethical framework including intelligent artefacts. *AISB/IACAP World Congress 2012 - The Machine Question: AI, Ethics and Moral Responsibility*. AISB, pp. 73-77.

Opaque Robots …

Poor Transparency -> robots that can **mislead** us[4]
-> choose to **trust**, or **lose confidence**

Robot behaviour **intends**
to mislead -> Robot **deceives** user.
Designer / owner responsible
for deception

Robot behaviour **unintentionally**
misleads -> **Failure** of designer / owner

4. P. a. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, 'A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction', Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5), 517–527, (2011).

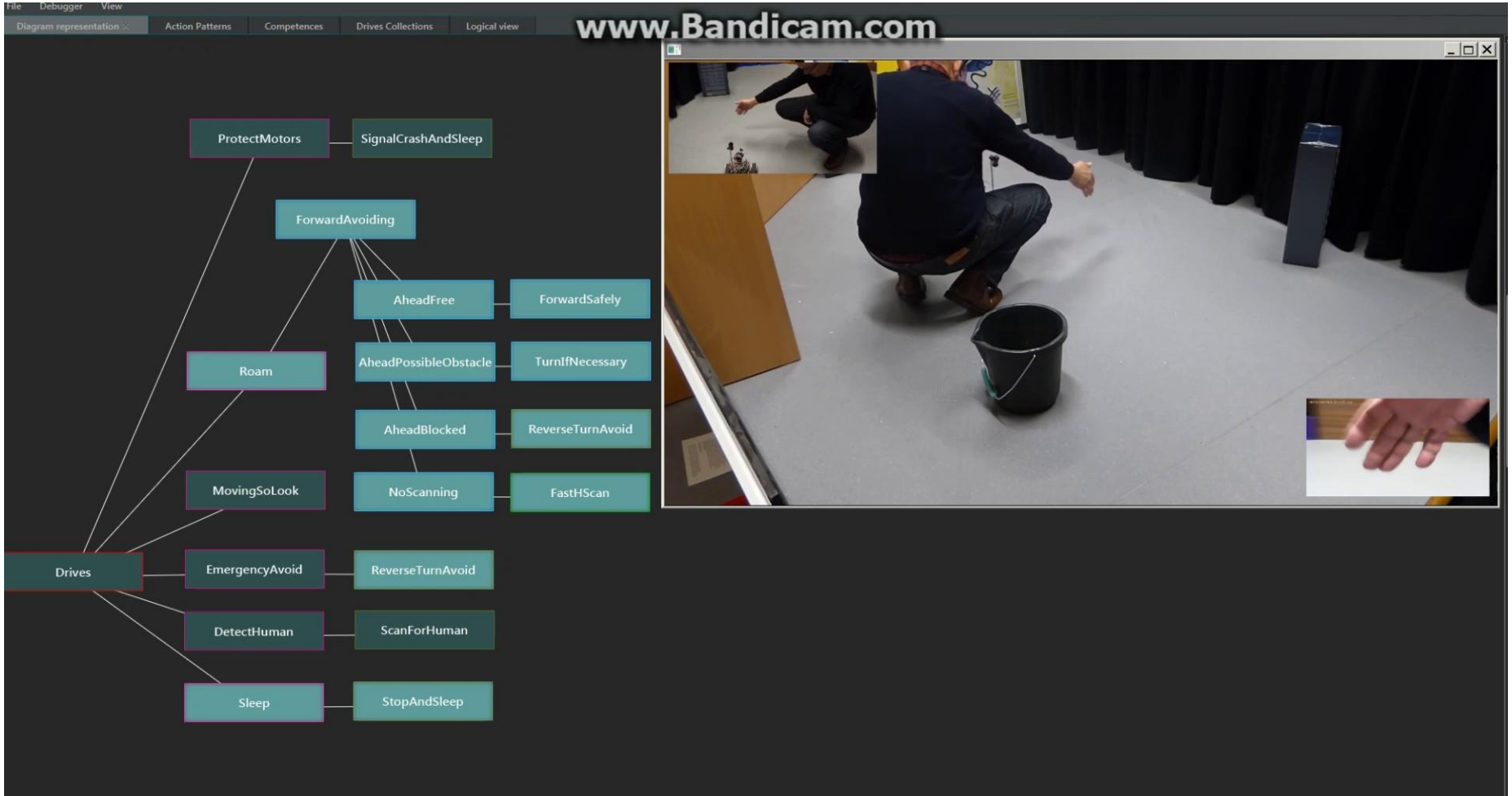
Research Questions

Is the **emotional impact of robots altered** by **understanding** their intelligence?

Can we build robots that **engage us emotionally, yet are transparent** in the way they interact with us?

Can we **build transparency into the substrate** of the machine architecture, such that it is an **implicit**, rather than explicit, **feature of the robot**?

Same video plus Transparency Display (ABOD3)



Post Treatment Questions:

Is the robot **thinking**? Y/N

Is the robot **intelligent**? 1-5

Can you tell what the robot is doing? Y/N

Describe robot task? Free text

Why do lights flash? Free text

What is person doing? Free text

Scored 0-2 for analysis

Significant Results (N=45)

Result	**Group One**	**Group Two**
Robot is thinking (0/1)	**0.36 (sd=0.48)**	**0.65 (sd=0.48)**
Robot Intelligence (1-5)	2.64 (sd=0.88)	2.74 (sd=1.07)
Understand objective (0/1)	0.68 (sd=0.47)	0.74 (sd=0.44)
Mental Model Accuracy (0-6)	**1.86 (sd=1.42)**	**3.39 (sd=2.08)**

1. **Marked difference** in the participants' **mental model accuracy** scores t(43)=2.86, p=0.0065, d=0.53
2. **No significant difference** in perceived **intelligence** between the two t(43)=0.35, p=0.73, d=0.29
3. **A substantially higher number** of participants in Group Two report that they **believe the robot is thinking**; t(43)=2.02, p=0.050

Conclusions from this Initial Study

1. Subjects can show **marked improvement in the accuracy of their mental model** of a robot observed on video, if they also **see** an accompanying display of the robot's **real-time decision making**.

2. An **improved mental model** of the robot is associated with an **increased perception of a thinking machine**, even though there is no significant change in the level of perceived intelligence.

3. The relationship between the perception of intelligence and thinking is not straightforward.

Transparent Minds …

Kinds of Minds[5]

- Darwinian - hardwired behaviours (phenotypes).
- Skinnerian - ABC Learning – associationism, behaviourism, connectionism
- Popperian - “permits our hypotheses to die in our head”
- Gregorian - imports tools from the external cultural environment.

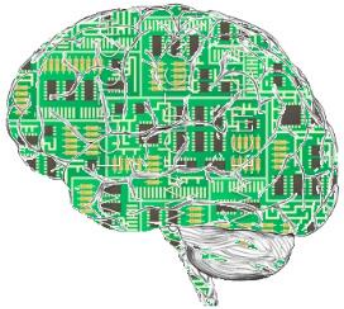
5. Kinds of Minds, Daniel Dennett, 1996, HarperCollins

Transparent Minds…

Human Mind

- D + S + P + G
- We are evolved -> share common abilities and goals.
- Theory of Mind
- Able to create narratives about own actions and those of others.

Robot Mind



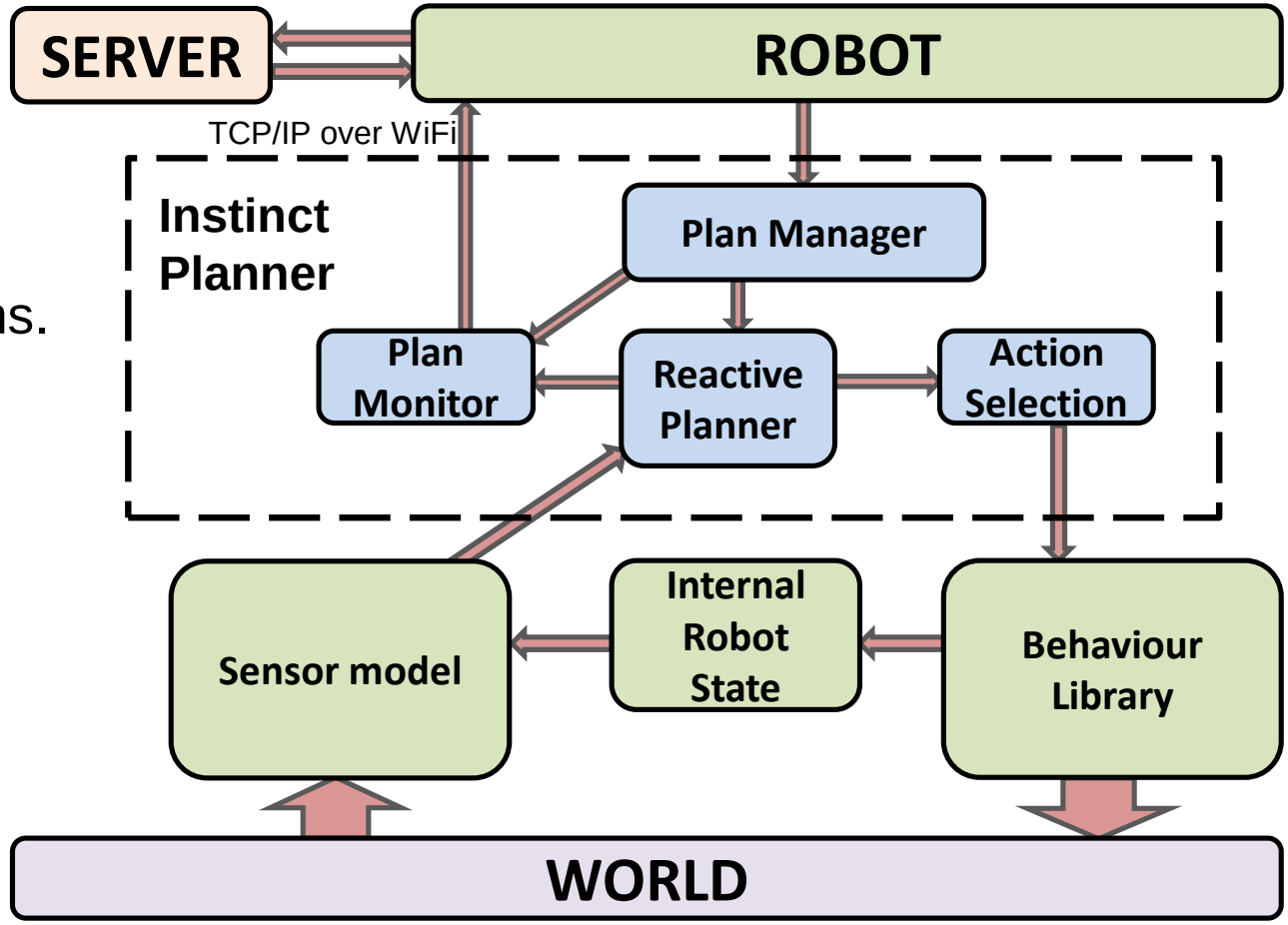
- D ¦ D + S ¦ ?P
- Designed not evolved
- No theory of mind of others
- Narrative meaning explicitly coded

5. Kinds of Minds, Daniel Dennett, 1996, HarperCollins

A *Transparent* Darwinian Mind

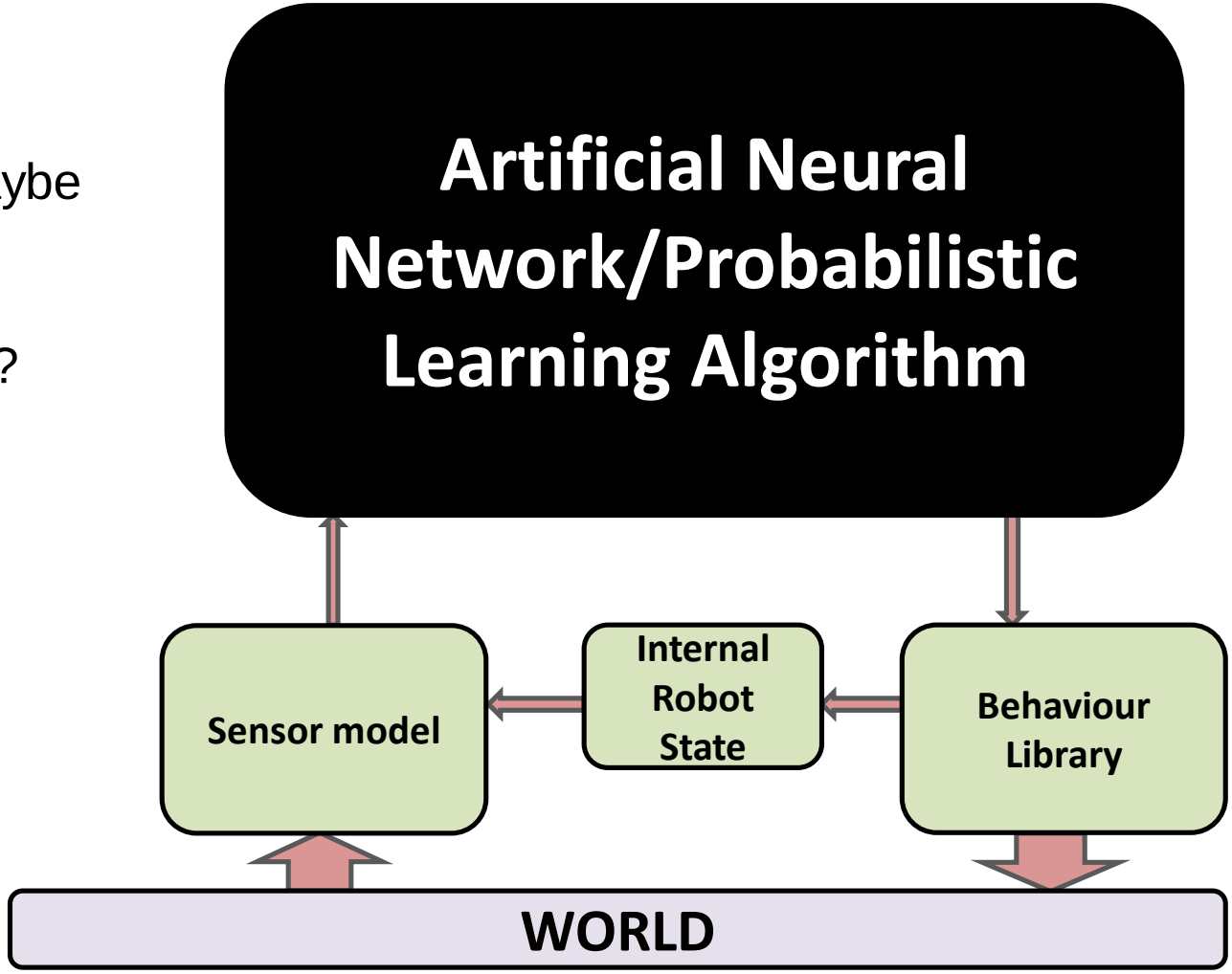
- Action Selection: Reactive Planner. Human readable plans.

- Narrative Generation From Monitoring:
  - Textual
  - Graphical
  - Verbal

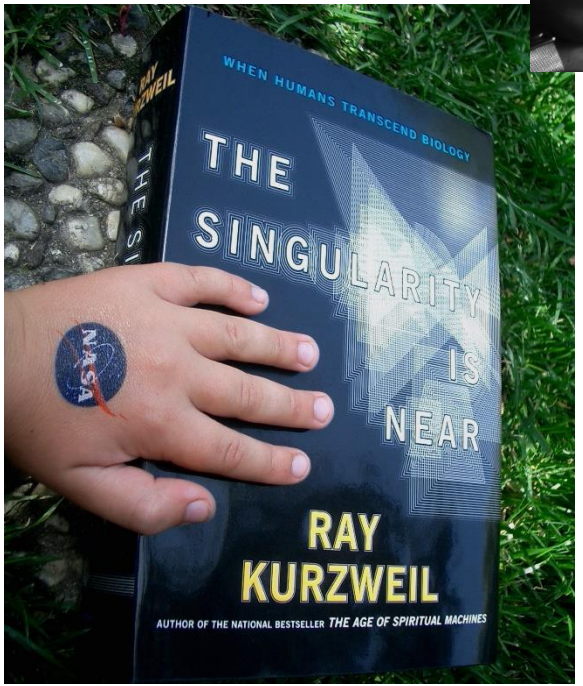


An *Opaque* Mind

- Mind - Darwinian, maybe Skinnerian
- Narrative Generation? **None**



?



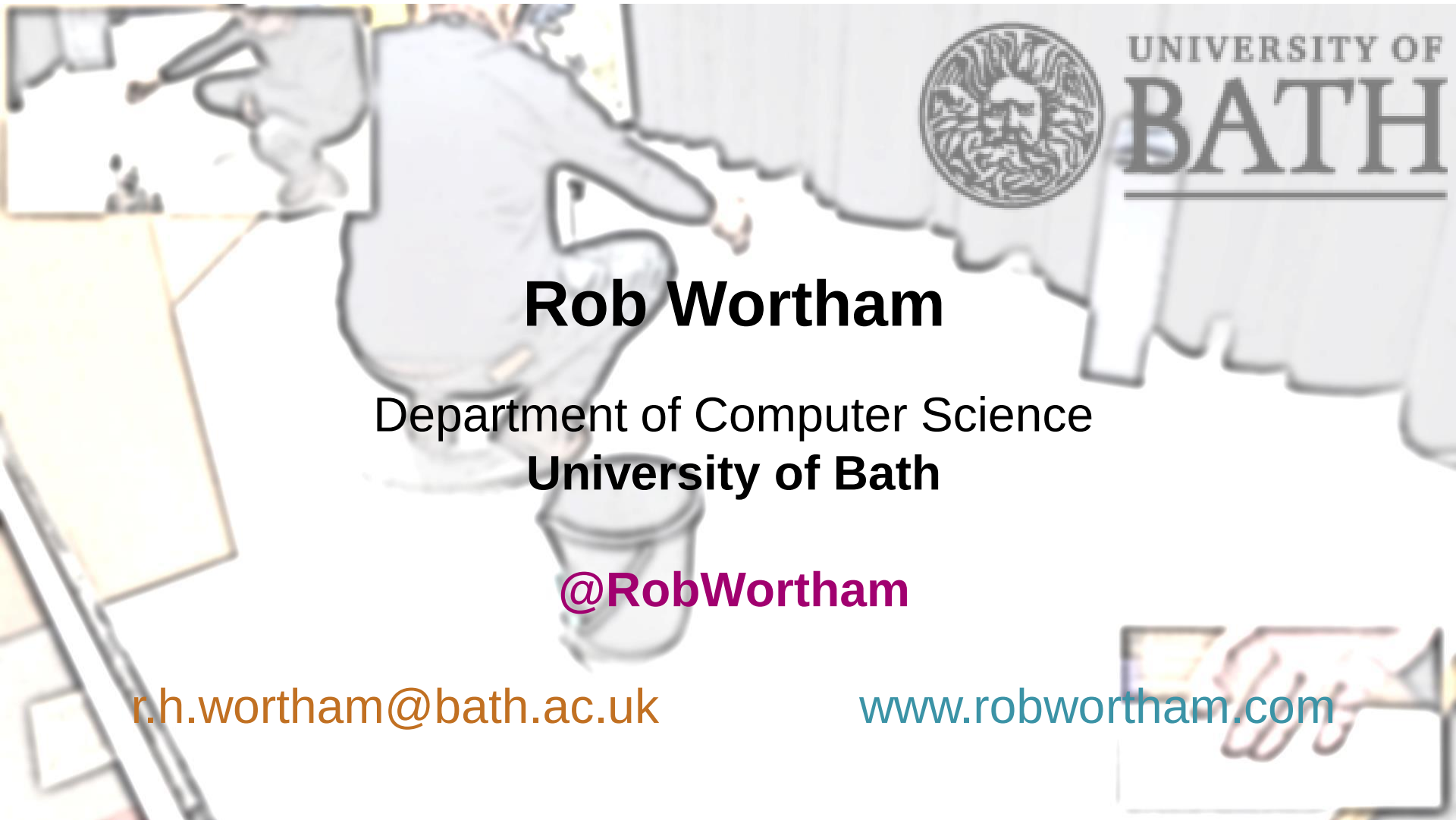
Artificial Neural Network/Probabilistic Learning Algorithm

WORLD

In Summary

1. There are **guidelines** for robot designers – the **EPSRC Principles**
2. **Robot transparency -> Improved mental model** -> **Increased perception of a thinking machine**. The relationship between the perception of intelligence and thinking is not straightforward.
3. We can build **transparent Darwinian minds** using reactive planning.
4. Transparency for ANN/Probabilistic approaches is a hard open research question (rule extraction).

Q: For action selection, is it better to focus on building ontogenetic (within lifetime) learning with reactive planning/other traditional approaches rather than black box approaches?



Rob Wortham

Department of Computer Science
University of Bath

@RobWortham

r.h.wortham@bath.ac.uk

www.robwortham.com