

What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems

Robert H. Wortham, Andreas Theodorou, Joanna J. Bryson

Department of Computer Science, University of Bath, UK

r.h.wortham@bath.ac.uk, a.theodorou@bath.ac.uk, jjb@alum.mit.edu

Abstract

Deciphering the behaviour of intelligent others is a fundamental characteristic of our own intelligence. As we interact with complex intelligent artefacts, humans inevitably construct mental models to understand and predict their behaviour. If these models are incorrect or inadequate, we run the risk of self deception or even harm. This paper reports progress on a programme of work investigating approaches for implementing robot transparency, and the effects of these approaches on utility, trust and the perception of agency. Preliminary findings indicate that building transparency into robot action selection can help users build a more accurate understanding of the robot.

1 INTRODUCTION

Article four of the EPSRC Principles of Robotics asserts that “Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.” [1]. Why is Transparency important, and how does it impact AI system design? Writers such as Mueller [10] suggest that as intelligent systems become both increasingly complex and ubiquitous, it becomes increasingly important that they are self explanatory, so that users can be confident about what these systems are doing and why. Mueller sees explanation as one of the three main characteristics of transparent computers, the others being dialogue and learning.

Humans have a natural if limited ability to understand others, however this ability has evolved and developed in the environment of human and other animal agency, which may make assumptions artificial intelligence does not conform to. Therefore it is the responsibility of the designers of intelligent systems to make their products transparent to us [17; 14]. This is of particular importance when deploying robots in environments where those who interact with them may be vulnerable, such as in care homes or hospitals [13], or equally in high-risk environments where misunderstanding a robot may have dangerous consequences.

Decoding the behaviour of intelligent others is a fundamental characteristic of our own intelligence. It is generally thought that many forms of effective interaction, whether

coercion or co-operation, rely on each party having some theory-of-mind (ToM) of the other [15; 12]. Individual actions and more complex behaviour patterns are thus interpreted within a pre-existing ToM framework. Whether that ToM is entirely accurate is unimportant, provided that it is predictive in terms of behaviour. Humans have a strong tendency to anthropomorphise not only nature, but anything around them [6] — the Social Brain Hypothesis [7] may explain this phenomenon. As we interact with complex intelligent artefacts, we construct anthropomorphic models to understand and predict their behaviour. If these models are incorrect, or inadequate, we are at best at risk of being deceived and at worse at risk of being harmed.

This paper reports preliminary findings from human-subject experiments concerning the understanding of a simple autonomous robot. Subjects watch a video of a robot interacting with a researcher, and report their theories about what the robot is doing and why. Some of these reports are wildly inaccurate, and interestingly many conclude that the robot’s objectives and abilities are far more complex than they in fact are. Importantly, we find that simply showing the runtime activation of the robot’s action selection along with the video results in users building significantly more accurate models of the robot’s behaviour.

2 BACKGROUND: REACTIVE PLANNING & ROBOT TRANSPARENCY

Here we use reactive planning techniques to build transparent autonomous agents. We have deployed the *Instinct Reactive Planner* [16] as the core action selection mechanism for the R5 robot. Instinct is deployed in the context of Bryson’s Behaviour Oriented Design (BOD) development methodology, as a replacement and extension of Parallel-rooted, Ordered Slip-stack Hierarchical (POSH) action selection¹ [4]. Instinct includes several enhancements taken from more recent papers extending POSH [11; 8], together with some ideas from other related planning approaches, notably Behaviour Trees (BT) [9]. A POSH plan consists of a *Drive Collection (DC)* containing one or more *Drives*. Each *Drive (D)* has a priority and a releaser. When the *Drive* is released as a result of sensory input, a hierarchical plan of *Competences, Action Patterns* and *Actions* follows.

¹POSH — <http://www.cs.bath.ac.uk/jjb/web/posh.html>

The Instinct planner has been specifically designed for low power processors and has a tiny memory footprint. Written in C++, it runs efficiently on both ARDUINO (ATMEL AVR) and MICROSOFT VC++ environments and has been deployed within a low-cost ARDUINO-based maker robot to study AI transparency. We have named this robot R5, in reference to the Rover 5 tracked platform on which it is based. Plans may be authored using a variety of tools including a promising visual design language *iVDL*, currently implemented using the DIA drawing package. The Instinct Planner and *iVDL* will shortly be made available on an open source basis².

2.1 Robot Plans

POSH plans are written in a LISP like notation, either using a text editor, or the Advanced Behaviour Oriented Design Environment (ABODE)³ editor [2], which allows graphical representations of the plans.

However, Instinct plans are written very differently, because they must use a much more compact notation due to memory constraints. We make use of the *Instinct Visual Design Language (iVDL)* — a graphical method of designing reactive plans, based on the ubiquitous Unified Modelling Language (UML) notation. UML is supported by many drawing packages and a simple PYTHON export script provided as part of Instinct allows plans to be created graphically within the DIA⁴ drawing tool. An example robot plan is shown in Figure 1. At this level of zoom the element details are not legible, but this screen shot gives an impression of how plans can be laid out, and the complexity of the plan used in our experiment.

2.2 The Transparent Planner

The Instinct Planner includes significant capabilities to facilitate plan design and runtime debugging. It reports the execution and status of every plan element in real time, allowing us to implicitly capture the reasoning process within the robot that gives rise to its behaviour. The planner has the ability to report its activity as it runs, by means of callback functions to a monitor class. There are six separate callbacks monitoring the Execution, Success, Failure, Error and In-Progress status events, and the Sense activity of each plan element. In the R5 robot, the callbacks write textual data to a TCP/IP stream over a wireless (wifi) link. A JAVA based Instinct Server receives this information and logs the data to disk. This communication channel also allows for commands to be sent to the robot while it is running. Figure 2 shows the overall architecture of the planner within the R5 robot, communicating via wifi to the logging server.

2.3 Realtime Plan Debugger

We are fortunate to have access to a new pre-alpha version of the ABODE plan editor, *ABOD3*, as seen in Figure 3. This version directly reads Instinct plans, and also reads a log file containing the real-time transparency data emanating from the Instinct Planner, in order to provide a real-time graphical display of plan execution.

²<http://www.robwortham.com/instinct-planner/>

³<http://www.cs.bath.ac.uk/~jjb/web/BOD/abode.html>

⁴Dia — <http://dia-installer.de/>

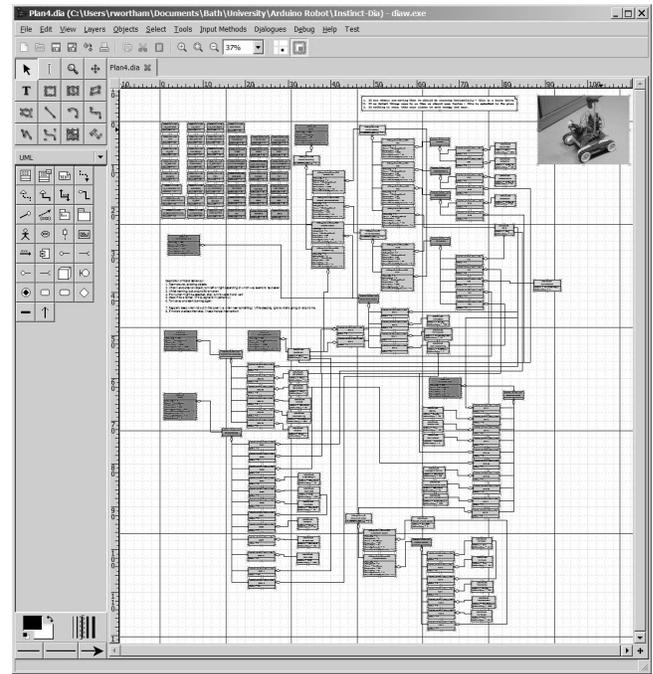


Figure 1: The Robot Plan Developed in *iVDL*

ABOD3 is also able to display a video and synchronise it with the debug display. In this way we are beginning to explore both runtime debugging and wider issues of AI Transparency.

3 METHODS: THE ROBOT EXPERIMENT

The robot in the video runs within an enclosed environment where it interacts with various objects and walls made of different materials. A researcher also interacts with the robot. The robot's action selection governs the behaviour of the robot by applying the reactive plan. A reactive plan encodes the priorities of a robot and the conditions when actions can be applied. A record of transparency data in the form of a log of which plan components are triggered at what time is collected by a remote server running on a laptop PC via a wifi connection.

Using its built-in real time clock the robot tags the transparency datastream with the start time of the experiment. It also includes the elapsed time in milliseconds with every datastream event. In this way the ABOD3 debugger is able to subsequently synchronise the datastream with video recordings taken during the experiment.

3.1 Robot Drives and Behaviours

The robot plan shown in figure 1 has six Drives. These are (in order of highest priority first):

- Sleep — this Drive has a ramping priority. Initially the priority is very low but it increases linearly over time until the Drive is released and completes successfully. The Drive is only released when the robot is close to an obstacle. This is to prevent the robot sleeping in the middle

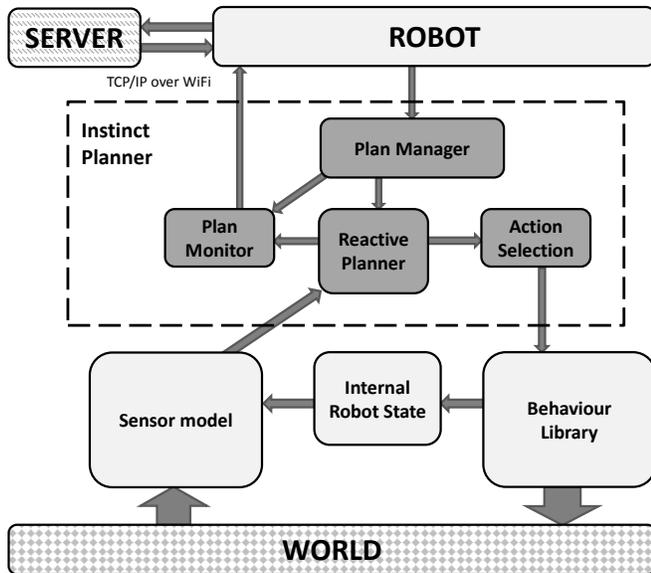


Figure 2: R5 Robot Software Architecture

of an open space where it may present a trip hazard. The sleep behaviour simply shuts down the robot for a fixed interval to conserve battery power.

- **Protect Motors** — released when the current drawn by the drive motors reaches a threshold. This might happen if the robot encounters a very steep incline or becomes jammed somehow. The Drive invokes an Action Pattern that stops the robot, signals for help and then pauses to await assistance.
- **Moving So Look** — simply enforces that if the robot is moving, it should be scanning ahead for obstacles.
- **Detect Human** — released when the robot has moved a certain distance from its last confirmed detection of a human, is within a certain distance of an obstacle ahead, and its Passive Infrared (PIR) detects heat that could be from a human. This Drive initiates a fairly complex behaviour of movement and coloured lights designed to encourage a human to move around in front of the robot. This continues to activate the PIR sensor thus confirming the presence of a human (or animal). It is of course not a particularly accurate method of human detection.
- **Emergency Avoid** — released when the robot's active infrared corner sensors detect reflected infrared light from a near obstacle. This invokes a behaviour that reverses the robot a small distance and turns left or right a fixed number of degrees. Whether to turn left or right is determined by which direction appears to be less blocked, as sensed by the active infrared detectors.
- **Roam** — released whenever the robot is not sleeping. It uses the scanning ultrasonic detector to determine when there may be obstacles ahead and turns appropriately to avoid them. It also modulates the robot speed and the

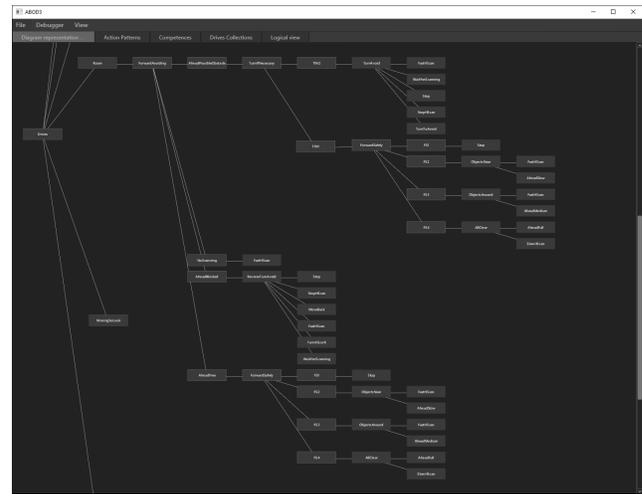


Figure 3: ABOD3 Showing Part of the Instinct Plan

rate of scanning depending on the proximity of obstacles.

3.2 Robot Videos

In order to quickly generate pilot initial results, we chose to video the robot rather than have participants interact with the robot directly. This research method has recently been chosen by others [5] with good results. Video has the benefit of ensuring all subjects share identical stimuli.



Figure 4: Video of interaction with the robot with no plan visible (stimulus for Group One).

The interaction is recorded from two positions at each end of the robot pen, and a camera mounted on a post attached to the robot also captures a 'robot's eye' view, providing a third perspective. The resulting composite video is approximately 5 minutes long. Figure 4 is a single frame from the video. It shows the researcher interacting with the robot. This video was shown to half of our group of test participants.

We used an early beta version of the ABOD3 tool to create a second video. A frame from this video is shown in figure 5. The six Drives described above are clearly visible. As each Drive is released and the associated behaviours are executed, the plan elements constituting the behaviours are highlighted.

Two reporting 2.74 (sd=1.07); $t(43)=0.35$ $p=0.73$ $d=0.29$. However, a substantially higher number of participants in Group Two (ABOD3) report that they believe the robot is thinking; $t(43)=2.02$, $p=0.050$.

4.2 Qualitative Outcomes

The data indicate very little emotional response to the robot in either group, with most participants indicating either ‘No Feeling’, and/or ‘Curious’. This data conflicts with the anecdotal feedback we frequently receive about the robot when people encounter it in our laboratory or during demonstrations. In these situations we often hear both quite strong positive & negative emotional reactions. Some find the robot scary or creepy, whilst others remark that it is cute, particularly when it is operational. At this stage it would appear that the remote nature of the video, and the small size of the robot on screen reduce the possibility of any significant emotional response.

From the answers to the question ‘why does the robot stop every so often’ it appears that this question is ambiguous. Some understand this to mean every time the robot stops to scan its environment before proceeding, and only one person took this to mean the sleep behaviour of the robot that results in a more prolonged period of inactivity. The question was intended to refer to the latter, and was particularly included because the Sleep Drive is highlighted by ABOD3 each time the robot is motionless with no lights flashing. However only one member of Group Two identified this from the video. Due to this ambiguity, the data related to this question was not considered further.

Many in Group Two, even those with a Science, Technology Engineering or Maths (STEM) degree level education, often still form a poor mental model of the robot. Here are some notable quotes from Group Two STEM participants that highlight this point:

- [the robot is] 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.
- [the robot is] 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?
- [the robot] is circling the room, gathering information about it with a sensor. It moves the sensor every so often in different parts of the room, so I think it is trying to gather spacial information about the room (its layout or its dimensions maybe).

5 DISCUSSION

There is a significant correlation between the accuracy of the participants’ mental models of the robot, and the provision of the additional transparency data provided by ABOD3. We have shown that a real-time display of a robots decision making produces significantly better understanding of that robot’s intelligence, even though that understanding may still include wildly inaccurate overestimation of the robot’s abilities.

Strikingly, there was one further significant result besides the improved mental model. Subjects who observed the real time display did not think the robot was more intelligent, but *did* think it ‘thought’ more. This result is counter-intuitive. We had expected that if ABOD3 resulted in increased transparency, that there would be a corresponding reduction in the use of anthropomorphic cognitive descriptions. However at least in this case the data suggests the reverse is true. When taken with the significant improvement in understanding of the robot’s actual drives and behaviours, this result implies that an improved mental model is associated with an increased perception of a thinking agent. Most likely this reflects the still pervasive belief that navigating in the real world is not a difficult task. *Intelligence* is a term that in ordinary language is often reserved for the things we apply conscious decision making to. The complexity of navigation, balance and perception has been persistently under-recognised [3].

Notably, while subjects exposed to the ABOD3 visualisations of the robot’s decision making considered the robot to be thinking more, they did not consider it to be more intelligent. In fact, the middling marks for intelligence in either condition may reflect a society-wide lack of certainty about the definition of the term rather than any cognitive assessment. Indeed the relatively large standard deviations for intelligence in Table 2 provide some evidence of this uncertainty.

It may be that the timing of the ABOD3 plan element highlighting & decay must be very tightly correlated with the video in order for them to appear connected. This may be similar to the temporal limit of lip synchronisation between video and audio when observing speech, beyond which the two data streams appear unrelated.

It may also be that other forms of transparency display would better serve non specialists i.e. those not familiar with reactive planning or the ABOD3 presentation paradigm.

The results also imply that 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. The relationship between the perception of intelligence and thinking is therefore not straightforward.

6 CONCLUSION & FURTHER WORK

We have demonstrated that 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. Although these are only the preliminary results of a small pilot study ($N = 45$), the outcome was strongly significant. The addition of ABOD3 visualisation of the robot’s intelligence does indeed make the machine nature of the robot more transparent.

There is clearly further work to be done to unpack the relationship between the improved mental model of the robot and the increased perception of a thinking machine.

This research confirms that the approach of using online video with web based questionnaires is both effective and efficient in terms of researcher time, and it has enabled us to quickly gather preliminary results from which further experi-

ments can be planned.

However, due to the lack of physical robot presence, it was not possible to gather any useful data about the emotional response of the participants. In situations where the emotional engagement of users to robots is of interest, the use of video techniques is therefore not recommended.

The technology used to construct the experimental system was found to be reliable, robust and straightforward to use. Given the low cost of the platform, we would recommend its use for similar low cost research robot applications.

The Instinct Planner combined with the iVDL graphical design tool enabled us to quickly generate a reliable yet sufficiently complex reactive plan for the R5 robot to allow us to conduct this experiment.

Despite using the early pre-alpha version of ABOD3, it confirmed its usefulness both as a tool during robot plan debugging, and to provide transparency information to untrained observers of the the robot.

References

- [1] M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden, T. Sorell, M. Wallis, B. Whitby, and A. Winfield. Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), April 2011. web publication.
- [2] C. Brom, J. Gemrot, M. Bida, O. Burkert, S. J. Partington, and J. J. Bryson. POSH Tools for Game Agent Development by Students and Non-Programmers. pages 1–8, 2006.
- [3] R. A. Brooks. Intelligence without reason. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, Sydney, Aug. 1991.
- [4] J. J. Bryson. Intelligence by design: principles of modularity and coordination for engineering complex adaptive agents. 2001.
- [5] D. Cameron, E. C. Collins, A. Chua, S. Fernando, O. McAree, U. Martinez-Hernandez, J. M. Aitken, L. Boorman, and J. Law. Help! i Can't reach the buttons: Facilitating helping behaviors towards robots. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9222, pages 354–358, 2015.
- [6] K. Dautenhahn. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, 4(1 SPEC. ISS.):103–108, 2007.
- [7] R. I. M. Dunbar. The Social Brain Hypothesis. *Evolutionary Anthropology*, pages 178–190, 1998.
- [8] S. Gaudl and J. J. Bryson. The Extended Ramp Goal Module: Low-Cost Behaviour Arbitration for Real-Time Controllers based on Biological Models of Dopamine Cells. *Computational Intelligence in Games 2014*, 2014.
- [9] C. U. Lim, R. Baumgarten, and S. Colton. Evolving behaviour trees for the commercial game DEFCON. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6024 LNCS(PART 1):100–110, 2010.
- [10] E. T. Mueller. *Transparent Computers: Designing Understandable Intelligent Systems*. Erik T. Mueller, San Bernardino, CA, 2016.
- [11] P. Rohlfshagen and J. J. Bryson. Flexible Latching: A Biologically-Inspired Mechanism for Improving the Management of Homeostatic Goals. *Cognitive Computation*, 2(3):230–241, 2010.
- [12] R. Saxe, L. E. Schulz, and Y. V. Jiang. Reading minds versus following rules: dissociating theory of mind and executive control in the brain. *Social neuroscience*, 1(3-4):284–98, jan 2006.
- [13] A. Sharkey and N. Sharkey. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, 2012.
- [14] A. Theodorou, R. H. Wortham, and J. J. Bryson. Why is my robot behaving like that ? Designing transparency for real time inspection of autonomous robots. In *EP-SRC Principles of Robotics Workshop, Proceedings of the AISB 2016 Annual Conference {accepted for publication}*, Sheffield, UK, 2016.
- [15] R. H. Wortham and J. J. Bryson. Communication. In *Handbook of Living Machines {accepted for publication}*. Oxford University Press, Oxford, 2016.
- [16] R. H. Wortham, S. E. Gaudl, and J. J. Bryson. Instinct : A Biologically Inspired Reactive Planner for Embedded Environments. In *Proceedings of ICAPS 2016 PlanRob Workshop {accepted for publication}*, London, UK, 2016.
- [17] R. H. Wortham, A. Theodorou, and J. J. Bryson. Robot Transparency, Trust and Utility. In *EP-SRC Principles of Robotics Workshop, Proceedings of the AISB 2016 Annual Conference {accepted for publication}*, Sheffield, UK, 2016.