# Improving Robot Transparency: Real-Time Visualisation of Robot AI Substantially Improves Understanding in Naive Observers

Robert H. Wortham[1], Andreas Theodorou[2] and Joanna J. Bryson[3]

*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. Here we demonstrate that providing even a simple, abstracted real-time visualisation of a robot's AI can radically improve the transparency of machine cognition. Findings from both an online experiment using a video recording of a robot, and from direct observation of a robot show substantial improvements in observers' understanding of the robot's behaviour. Unexpectedly, this improved understanding was correlated in one condition with an increased perception that the robot was 'thinking', but in no conditions was the robot's assessed intelligence impacted. In addition to our results, we describe our approach, tools used, implications, and potential future research directions.**

## I. INTRODUCTION

The fourth of the five 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? There has been considerable previous research to investigate ways in which robots can understand humans [2]. However transparency is the converse. Here we are interested in how robots should be designed in order that we can understand them.

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 necessarily conform to. Therefore it is the responsibility of the designers of intelligent systems to make their products transparent to us [3][4].

It is generally thought that many forms of effective interaction, whether cooperative or coercive, rely on each party having some theory of mind (ToM) concerning the other [5][6]. Individual actions and complex behaviour patterns can be more easily interpreted within a pre-existing ToM framework, often created through modelling from one's own expectations by projection to the others' identity. Whether that ToM is entirely accurate is unimportant, provided that it is sufficiently predictive to inform one's own action selection [7]. Ideally such 'good enough' modelling should include an accurate assessment of how inaccurate our model might be. However, in the case of AI humans have been repeatedly shown to over-identify with machines, even to their own detriment [8]. This holds true for 6-month-old babies, so cannot be attributed to or easily solved by implicit enculturation [9].

In robot-human collaborative scenarios, transparency has been shown to improve the quality of teamwork [10]. It is also a key factor when humans attribute credit and blame in these collaborative scenarios [11]. Increased robot transparency is associated with reduced assignment of credit or blame to the robot, and increased assignment to humans. This increased focus on and facilitation of human agency in collaborative robot-human tasks is a desirable outcome, because it allows automation to empower and enhance its human users.

Writers such as Mueller [12] and Cramer [13] 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. Robot designers have long recognised that any complex autonomous control strategy, combined with the complex real-world environment that differentiates robotics from ordinary AI, necessarily results in non-repeatable behaviour and unexpected conditions [14]. Whilst many authors have recently focussed on dialogue and explanation as a solution to transparency, such systems are not appropriate to every circumstance, both because of the computational overhead for AI natural language systems, and the cognitive and temporal costs of dialogue.

Note that the need for users to form a useful model of a robot is orthogonal to issues of verification of robot behaviour. Whilst others have concentrated their research on making a robot safe and predictable [15][16], here we are interested here in the models that observers of a robot use to understand and predict its behaviour. The novelty of our experiments is that unlike other transparency studies in the literature which concentrate on human-robotics collaboration, our study focuses on unplanned robot encounters, where human interactors were not necessarily anticipating working with an artificial system at all, let alone a particular system they may have been trained to use.

Here we demonstrate that even abstracted and unexplained real-time visualisation of a robot's priorities can substantially improve human understanding of machine intelligence, including for naive users. Subjects watch a video of, or directly

*The work of Andreas Theodorou was supported by EPSRC Grant EP/L016540/1.

[1]Robert H. Wortham is with Department of Computer Science, University of Bath, Bath, BA2 7AY, UK r.h.wortham@bath.ac.uk

[2]Andreas Theodorou is with the Department of Computer Science, University of Bath, Bath, BA2 7AY, UK a.theodorou@bath.ac.uk

[3]Joanna J. Bryson is with the Department of Computer Science, University of Bath, Bath, BA2 7AY, UK; and Princeton Center for Information Technology Policy, Princeton, NJ, 08544, USA. j.j.bryson@bath.ac.uk

observe, 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. Nevertheless and importantly, we find that simply showing the runtime activation of the robot's action selection along with the its behaviour results in users building significantly more accurate models. To our knowledge, this is the first real-time visual presentation of reactive robot plans using a graphical plan representation.

## II. TECHNOLOGIES USED: REACTIVE PLANNING & ROBOT TRANSPARENCY

Here we use reactive planning techniques to build transparent AI for an autonomous robot. We have deployed the *Instinct* reactive planner [17] as the core action selection mechanism for the R5 robot, shown in Figure 2. The R5 robot is named in reference to the Rover 5 tracked platform on which it is based. Instinct is deployed in the context of the Behaviour Oriented Design (BOD) development methodology, as a replacement and extension of Parallel-rooted, Ordered Slip-stack Hierarchical (POSH) action selection[1] [18]. Instinct includes several enhancements taken from recent papers extending POSH [19], [20], together with some ideas from other related planning approaches, notably Behaviour Trees (BT) [21]. A POSH plan consists of a *Drive Collection (DC)* containing one or more *Drives*, which can be thought of as possible goals for the system. Each Drive (D) has a priority and a releaser. When the *Drive* is released as a result of sensory input, a hierarchical reactive subplan follows, but this subplan can be interrupted before completion if the situation changes and a higher-priority goal is triggered.

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 for this study of AI transparency. Plans may be authored using a variety of tools including a visual design language *iVDL*, currently implemented using the DIA drawing package. The Instinct Planner and iVDL are available on an open source basis[2].

### A. 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 1 shows the overall architecture of the planner within the R5 robot, communicating via WiFi to the logging server.

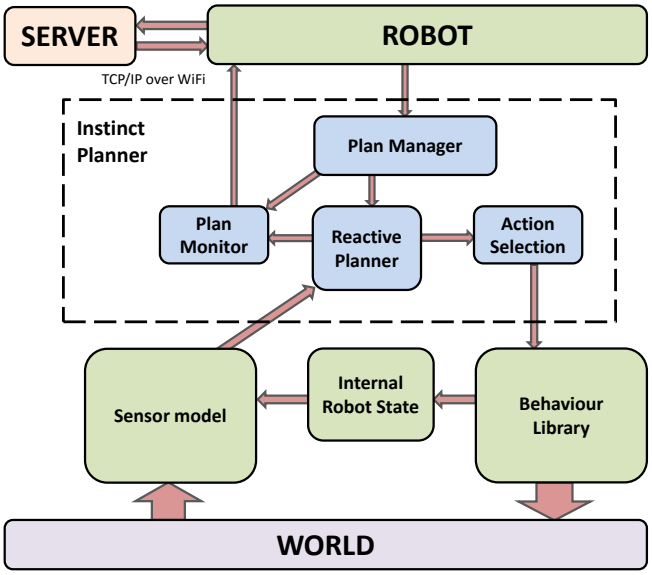


Fig. 1. R5 Robot Software Architecture. The arrows represent the primary data flows between the various modules.

### B. Robot Drives and Behaviours

The robot's overall function is to search a space looking for humans. Typical real world applications would be search and rescue after a building collapse, or monitoring of commercial cold stores or similar premises.

The robot reactive plan 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 and is inhibited whilst the robot confirms the presence of a human. This is to prevent the robot sleeping in the middle 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. This has a high priority so that this rule is always enforced whatever else the robot may be doing. However, it is only triggered when the robot is moving and the head is not scanning.

[1] http://www.cs.bath.ac.uk/~jjb/web/posh.html
[2] http://www.robwortham.com/instinct-planner/

- 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 nearby 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's speed and the rate of scanning depending on the proximity of obstacles.

### C. Real-Time Plan Debugger

We use the new version of the ABODE plan editor for POSH plans, *ABOD3*, as seen in Figure 3 [22]. This is the first version of ABODE to support real-time visualisation. We modified ABOD3 to directly read Instinct plans, and also to read 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. ABOD3 is also able to display a video and synchronise it with the debug display. In this way we can explore both runtime debugging and wider issues of AI Transparency. This facility is used in our first experiment. ABOD3 is also able to process and display a real-time feed of transparency data directly from the R5 robot as it runs. This facility is used in our second experiment.

## III. METHODS: THE ROBOT EXPERIMENTS

Two separate experiments are described. The first uses a video recording of the R5 robot and a web based online questionnaire. The second experiment involves participants directly observing the robot in a public space.

### A. Experiment One — Online Video

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. As mentioned earlier, a reactive plan encodes the (proactive) priorities of an autonomous 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.

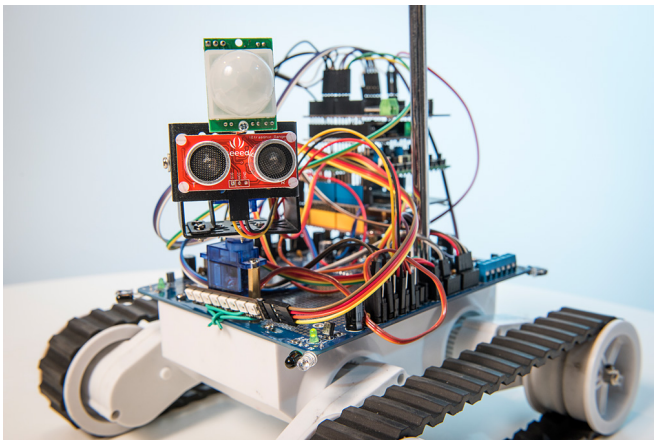

Fig. 2. The R5 Robot used in the experiments. This is a low cost tracked maker robot based on the Arduino platform. It has infrared proximity sensors and a moving 'head' with two degrees of freedom. The head carries ultrasonic range finding and a PIR sensor to detect humans. R5 uses wifi to send transparency data back to the ABOD3 realtime plan visualisation tool.

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.

Fig. 3. ABOD3 display of part of the Instinct plan described in the text. Note the element labels are readable on the original display.

*1) Robot Videos:* For our initial study, we chose to video the robot rather than have participants interact with the robot directly. This research method has recently been chosen by others [23] with good results. Video has the benefit of ensuring all subjects share identical stimuli.

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 five 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.

Using the ABOD3 tool, we created a second video. A

Fig. 4. Video of interaction with the robot with noc plan visible (stimulus for Group One in the first experiment).

frame from this video is shown in Figure 6. 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. This highlighting is synchronised with the behaviour of the robot visible in the video. This gives the viewer access to a great deal more information from the robot than is available by watching the robot alone. ABOD3 conveniently allows us to collapse the lower levels in the hierarchy, and position the visible plan elements for ease of understanding. For the purpose of clarity in the video, we chose to display only the highest levels of the reactive plan, primarily the Drives.

*2) Demographic & Post-Treatment Questionnaires:* For the Online video experiment, the participants were initially sent an email questionnaire to gather basic demographic data: age, gender, educational level, whether they use computers, whether they program computers and whether they have ever used a robot. Based on this information they were then divided into two groups that were matched as nearly as possible for participant mix. Each group received an identical email asking them to carefully watch a video and then answer a second questionnaire. Group One was directed to the composite video (Fig 4), and Group Two to the debug video (Fig 6).

TABLE I
POST-TREATMENT QUESTIONS

| Question | Response | Category |
|---|---|---|
| Is robot thinking? | Y/N | Intel |
| Is robot intelligent? | 1-5 | Intel |
| Feeling about robot? | Multi choice | Emo |
| Understand objective? | Y/N | MM |
| Describe robot task? | Free text | MM |
| Why does robot stop? | Free text | MM |
| Why do lights flash? | Free text | MM |
| What is person doing? | Free text | MM |
| Happy to be person? | Y/N | Emo |
| Want robot in home? | Y/N | Emo |

Table I summarises the questions asked after the participant had seen the video. These questions are designed to measure various factors: the measure of intelligence perceived by the participants (Intel), the emotional response (if any) to the robot (Emo), and—most importantly—the accuracy of the participants' mental model of the robot (MM). For analysis, the four free text responses were rated for accuracy with the robot's actual Drives & behaviours and given a score per question of 0 (inaccurate or no response), 1 (partially accurate) or 2 (accurate). The marking was carried out by a single researcher for consistency, without access to either subject identities or knowledge of which group the subject was in. No special vocabulary was expected. The questions used in the questionnaire are deliberately very general, so as not to steer the subject. Similarly, the marking scheme used was deliberately coarse grained because we are looking for a significant effect at the general level of understanding, not for a nuanced improvement in the subject's model. Question 3 was found to be ambiguous and so is not included in the scores, see below. By summing the scores, the accuracy of the participant's overall mental model is scored from 0 to 6.

### *B. Experiment Two — Directly Observed Robot*

This subsequent experiment took place over three days at the At-Bristol Science Learning Centre, Bristol, UK. This context was chosen because of available subjects in a controlled setting.

Fig. 5. The arrangement of the Directly Observed Robot experiment at At-Bristol. Obstacles visible include a yellow rubber duck and a blue bucket. The position and orientation of the transparency display is shown.

The robot operated within an enclosed pen as a special interactive exhibit within the main exhibition area, see Fig 5. Visitors, both adults and children, were invited to sit and observe the robot in operation for several minutes whilst the robot moved around the pen and interacted with the researchers. Subjects were expected to watch the robot for at least three minutes before being handed a paper questionnaire. They then completed the questionnaire, which contained the same questions as for the Online Video experiment above. During this time subjects were able to continue to watch the robot in operation.

A large computer monitor was positioned at the front of the pen displaying the ABOD3 real-time visualisation of plan execution. This display was either enabled or disabled for

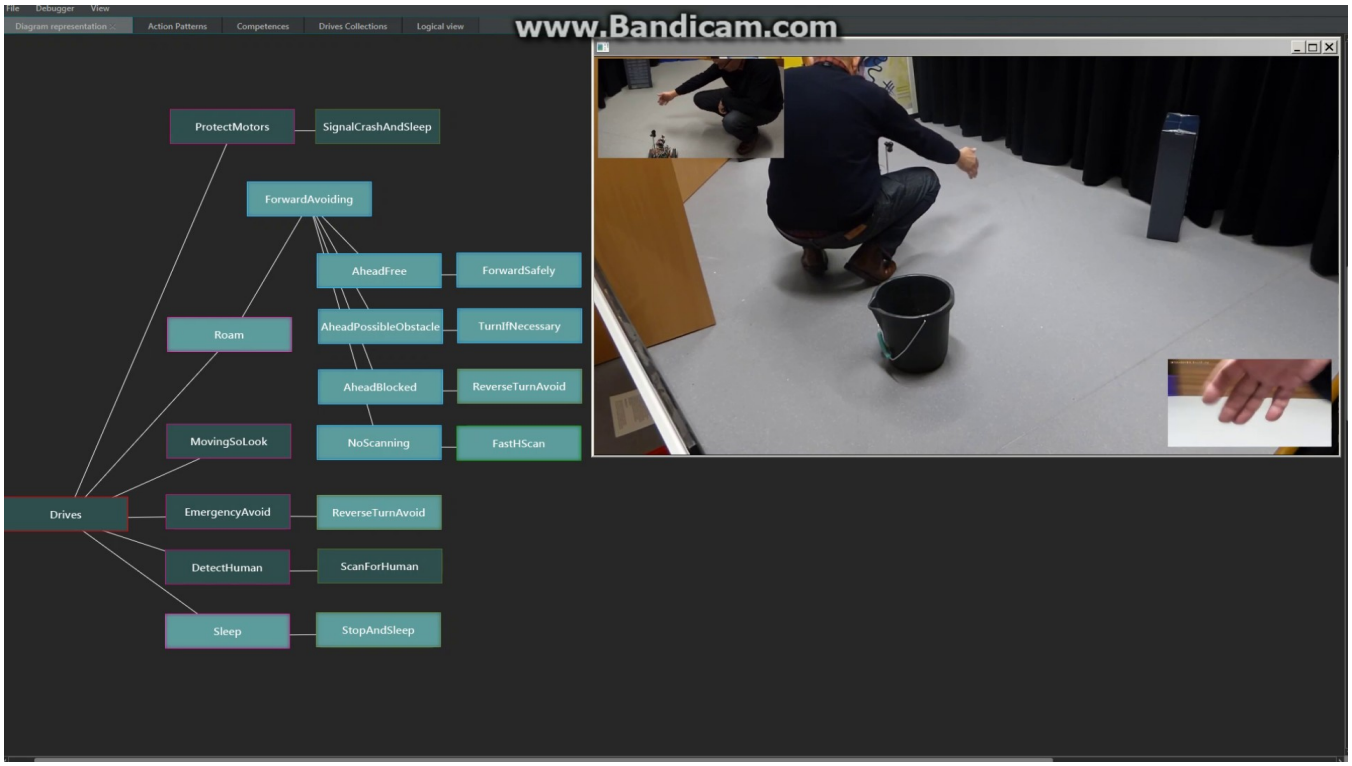


Fig. 6. More transparent video showing the ABOD3 plan representation; sub-trees have been hidden from view (stimulus for Group Two). Note that the drive labels were legible to the subjects, and can be seen clearly in the printed version of this paper, or by zooming the postscript version. The Drives are also explained above in Section II-B

periods as the days progressed to create the Group 2 and Group 1 datasets. Only adult data (age eighteen and above) is included in the results presented in this paper.

## IV. RESULTS

The demographics of each group of participants is shown in Table II and Table III. For the Online Video experiment it was possible to match the groups prior to watching the video. Priority was given to matching the number of programmers in each group, and to having an equal gender mix. This was not possible in the Directly Observed Robot experiment, however Table III shows the groups were nevertheless well-balanced.

TABLE II
ONLINE VIDEO EXPERIMENT: DEMOGRAPHICS OF PARTICIPANT GROUPS ($N = 45$)

| Demographic | Group One | Group Two |
|---|---|---|
| Total Participants | 22 | 23 |
| Mean Age (yrs) | 39.7 | 35.8 |
| Gender Male | 11 | 10 |
| Gender Female | 11 | 12 |
| Gender PNTS | 0 | 1 |
| STEM Degree | 7 | 8 |
| Other Degree | 13 | 13 |
| Ever worked with a robot? | 2 | 3 |
| Do you use computers? | 19 | 23 |
| Are you a Programmer? | 6 | 8 |

TABLE III
DIRECTLY OBSERVED ROBOT EXPERIMENT: DEMOGRAPHICS OF PARTICIPANT GROUPS ($N = 55$)

| Demographic | Group One | Group Two |
|---|---|---|
| Total Participants | 28 | 27 |
| Mean Age (yrs) | 48.0 | 40.0 |
| Gender Male | 10 | 10 |
| Gender Female | 18 | 17 |
| STEM Degree | 5 | 9 |
| Other Degree | 11 | 8 |
| Ever worked with a robot? | 7 | 6 |
| Do you use computers? | 20 | 22 |
| Are you a Programmer? | 6 | 5 |

### A. Main Findings

The primary results obtained from the experiments are outlined in Table IV and Table V. Data is analysed using the unpaired t test. First and most importantly, in both experiments there is a marked difference in the participants' mental model accuracy scores between Group One (just observe robot) and Group Two (observe robot and debug display). This confirms 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. Online Video experiment; t(43)=2.86, p=0.0065, Directly Observed Robot experiment t(55)=3.39, p=0.0013.

Secondly, there is no significant difference in perceived

TABLE IV

ONLINE VIDEO EXPERIMENT: MAIN RESULTS. BOLD FACE INDICATES RESULTS SIGNIFICANT TO AT LEAST $p = .05$.

| Result | Group One | Group Two |
|---|---|---|
| **Is thinking (0/1)** | 0.36 (sd=0.48) | 0.65 (sd=0.48) |
| 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) |
| **Report Accuracy (0-6)** | 1.86 (sd=1.42) | 3.39 (sd=2.08) |

TABLE V

DIRECTLY OBSERVED ROBOT EXPERIMENT: MAIN RESULTS. BOLD FACE INDICATES RESULTS SIGNIFICANT TO AT LEAST $p = .05$.

| Result | Group One | Group Two |
|---|---|---|
| Is thinking (0/1) | 0.46 (sd=0.50) | 0.56 (sd=0.50) |
| Intelligence (1-5) | 2.96 (sd=1.18) | 3.15 (sd=1.18) |
| **Understand objective (0/1)** | 0.50 (sd=0.50) | 0.89 (sd=0.31) |
| **Report Accuracy (0-6)** | 1.89 (sd=1.40) | 3.52 (sd=2.10) |

robot intelligence between the two groups in each experiment, although across experiments the data indicates a slightly higher level of perceived intelligence when the robot was directly observed.

Thirdly, in the Online Video experiment, 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. However, this effect is not significantly repeated when the robot is directly observed; t(55)=0.680, p=0.500.

Finally, for participants directly observing the robot, the ABOD3 display significantly affects their report that they understand what the robot is trying to do; t(55)=3.44, p=0.0011. This is not the case in the Online Video experiment, where the Group 2 data shows no significant affect; t(43)=0.425, p=0.673.

### B. Qualitative Outcomes

Participants also select from a list of possible emotional states: Happy, Sad, Scared, Angry, Curious, Excited, Bored, Anxious, No Feeling. For the Online Video experiment the data indicate very little emotional response to the robot in either group, with most participants indicating either 'No Feeling', or only 'Curious'. However, in the Directly Observed Robot experiment, participants indicate a higher level of emotional response, summarised in Table VI; t(98)=2.63, p=0.0098.

We had predicted the robot might be more emotionally salient when it was experienced directly. However, from Table VI it can be seen that curiosity dominates the results. Nevertheless, the addition of the transparency display may well increase the emotions reported; t(53)=1.91, p=0.0622. This may be a topic for future investigation.

In the first Online Video experiment, 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 in this dataset. This question was subsequently refined in the second, Directly Observed Robot experiment to 'Why does it just stop every so often (when all its lights go out)?'. Six participants then correctly answered this question and so it is included in the analysis.

TABLE VI

DIRECTLY OBSERVED ROBOT EXPERIMENT: SELF REPORTED EMOTION ($N = 55$)

| Reported Emotion | Group One | Group Two |
|---|---|---|
| Curious | 23 | 23 |
| Excited | 5 | 10 |
| Happy | 5 | 12 |
| No Feeling | 4 | 2 |
| Anxious | 0 | 1 |
| Bored | 1 | 0 |
| Scared | 1 | 0 |

Despite the improved performance of Group Two, many members, even those with a Science, Technology Engineering or Maths (STEM) degrees, still form a poor mental model of the robot. Here are some notable quotes from STEM participants:

- [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.* [we are unsure of the picture to which the participant refers]
- *is it trying to identify where the abstract picture is and how to show the complete picture?* [picture visible in Figure 4]
- [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).*
- [the robot] *maybe finding certain colours.*

These comments indicate that in the absence of an accurate model, environmental cues and possibly previous knowledge of robots are used to help create a plausible narrative.

## V. DISCUSSION

Across both experiments, 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 robot's decision making produces significantly better understanding of that robot's intelligence, even though

that understanding may still include wildly inaccurate over-estimation of the robot's abilities.

Strikingly, there was one further significant result besides the improved mental model. Subjects in Experiment 1 (Online Video) 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, so the amount of different planning steps employed by the robot during the process may come as a surprise. Notably, with the immediate presence of the robot in the shared environment in the second experiment, assessments of thinking under both conditions moved towards '50–50' or complete uncertainty, though the trend was still in the same direction.

Unlike *thinking*, *intelligence* seems to be a term that in ordinary language is often reserved for conscious decision making. Notably, even where 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. The relatively large standard deviations for intelligence in Tables IV and V provide some evidence of this uncertainty. Comparing results from the two experiments, it might be that the immediacy of the directly observed robot makes the objective more confusing without transparency and more apparent with transparency. Further investigation would be required to confirm whether this is repeatable.

In the first experiment, the lack of emotion with respect to the robot was unexpected, and conflicts with the spontaneous feedback we frequently receive about the R5 robot when people encounter it in our laboratory or during demonstrations. In these situations we often hear both quite strong positive and negative emotional reactions. Some find the robot scary or creepy [24], whilst others remark that it is cute, particularly when it is operational. We hypothesise that the remote nature of the video, or the small size of the robot on screen, reduce the chance of significant emotional response. Indeed this is confirmed by the higher levels of emotional response measured when participants directly observe the robot. Lack of creepyness (Anxious, Scared) may be due to the more controlled setting of the experiment, or the presence of 'experts' rather than peers. It is also interesting that the transparency display appears to further solicit positive emotional responses. Perhaps this reflects a delight or satisfaction that the robot behaviour is 'explained' by the display.

## VI. CONCLUSION & FURTHER WORK

We have demonstrated that subjects can show marked improvement in the accuracy of their mental model of a robot observed either directly or on video, if they also see an accompanying display of the robot's real-time decision making. In both our pilot study using online video ($N = 45$) and our subsequent experiment with direct observation ($N = 55$), 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.

The results of the Online Video experiment 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. However, this effect is not seen when the robot is directly observed. The relationship between the perception of intelligence and thinking is therefore not straightforward. 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.

Experiment 1 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 experiments can be planned. However, we did not gather any useful data about the emotional response of the participants using this methodology. This may possibly be due to the lack of physical robot presence. Therefore, in situations where the emotional engagement of users to robots is of interest, the use of video techniques may prove ineffective. We intend to explore this further in future work. We also intend to use the Godspeed questionnaire [25] in subsequent studies to facilitate comparison with the future work of others.

The technology used to construct the experimental system was found to be reliable, robust and straightforward to use. 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. The robot and real-time ABOD3 operated reliably over three days without problems despite some unexpected participant physical handling. Given the low cost of the platform, we would recommend its use for similar low cost research robot applications.

The fact that good results were achieved with a pre$-\alpha$ version of ABOD3 gives us high hopes for its utility not only for visualisation but also for real-time plan debugging. Certainly it proved able to provide transparency information to untrained observers of an autonomous robot.

This paper reports substantial significant impact of simply exposing the real time control state of a robot in two experiments involving lay subjects. A lay observer and technical specialist need different levels of detail. Future work could include varying the design of the visualisation dependent both on the robot task and user type.

## 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] M. K. Lee and M. Makatchev, "How Do People Talk with a Robot ? An Analysis of Human-Robot Dialogues in the Real World," *CHI 2009 Spotlight on Works in Progress*, pp. 3769–3774, 2009.

[3] R. H. Wortham and A. Theodorou, "Robot transparency, trust and utility," *Connection Science*, vol. 29, no. 3, pp. 242–248, 2017.

[4] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Designing and implementing transparency for real time inspection of autonomous robots," *Connection Science*, vol. 29, no. 3, pp. 230–241, 2017.

[5] R. H. Wortham and J. J. Bryson, "Communication," in *Living Machines {in press}*, T. J. Prescott and P. F. M. J. Verschure, Eds. Oxford: Oxford University Press, 2017.

[6] 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*, vol. 1, no. 3-4, pp. 284–98, jan 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18633794

[7] L. W. Barsalou, "Simulation, situated conceptualization, and prediction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1281–1289, 2009. [Online]. Available: http://rstb.royalsocietypublishing.org/content/364/1521/1281

[8] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. New York, NY, USA: ACM, 2015, pp. 141–148. [Online]. Available: http://doi.acm.org/10.1145/2696454.2696497

[9] K. Kamewari, M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki, "Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion," *Cognitive Development*, vol. 20, no. 2, pp. 303–320, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885201405000201

[10] C. Breazeal, C. Kidd, A. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Alberta, Canada: Ieee, 2005, pp. 708–713. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1545011

[11] T. Kim and P. Hinds, "Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction," *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 80–85, 2006.

[12] E. T. Mueller, *Transparent Computers: Designing Understandable Intelligent Systems*. San Bernardino, CA: Erik T. Mueller, 2016.

[13] H. S. M. Cramer, "Interaction with user-adaptive information filters." *CHI 2007 extended abstracts on Human factors in computing systems*, p. 1633, 2007.

[14] T. H. J. Collett and B. a. MacDonald, "Developer Oriented Visualisation of a Robot Program An Augmented Reality Approach," *HRI '06 Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 49–56, 2006.

[15] M. Fisher, L. Dennis, and M. Webster, "Verifying Autonomous Systems," *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013. [Online]. Available: http://doi.acm.org/10.1145/2494558

[16] A. F. T. Winfield, C. Blum, and W. Liu, "Towards an Ethical Robot : Internal Models , Consequences and Ethical Action Selection," in *Advances in Autonomous Robotics Systems*, M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, Eds. Switzerland: Springer International Publishing, 2014, pp. 85–96. [Online]. Available: http://link.springer.com/content/pdf/10.1007%2F978-3-319-10401-0_8.pdf

[17] 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*, London, UK, 2016. [Online]. Available: http://icaps16.icaps-conference.org/proceedings/planrob16.pdf

[18] J. J. Bryson, "Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents," Ph.D. dissertation, MIT, 2001. [Online]. Available: ftp://ll1.ai.mit.edu/people/joanna/phd-tr.pdf

[19] P. Rohlfshagen and J. J. Bryson, "Flexible Latching: A Biologically-Inspired Mechanism for Improving the Management of Homeostatic Goals," *Cognitive Computation*, vol. 2, no. 3, pp. 230–241, 2010.

[20] 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. [Online]. Available: http://opus.bath.ac.uk/40056/

[21] 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)*, vol. 6024 LNCS, no. PART 1, pp. 100–110, 2010.

[22] A. Theodorou, "Abod3: A graphical visualization and real-time debugging tool for bod agents," in *EUCognition 2016*, Vienna, Austria, December 2016. [Online]. Available: http://opus.bath.ac.uk/53506/

[23] 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)*, vol. 9222, 2015, pp. 354–358.

[24] F. T. McAndrew and S. S. Koehnke, "On the nature of creepiness," *New Ideas in Psychology*, vol. 43, pp. 10–15, 2016.

[25] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.